

## Materials and Methods

### Multiple sequence alignment generation

Initial gene sequence alignments were generated using novel software as follows. Starting from sequences in the seed genome of *Cyanidioschyzon merolae* (which was chosen because of its large well annotated proteome) a blast analysis was conducted to identify putatively homologous sequences in each of the additional eukaryotic genomes. Selection of sequences was based on the BLAST 'expect' score ( $<1e-4$ ), and maximising the total coverage ( $>60\%$ ) and similarity ( $>40\%$ ) based on the 'high-scoring sequence pairs' (HSP's). In addition, a reciprocal BLAST search was conducted with the best scoring sequence against the seed genome, and the sequence selected only if the top scoring hit was the original seed sequence. When all eukaryotic sequences had been identified, they were in-turn used as seed sequences to identify a most likely homologous sequence within the archaeabacterial genomes. This sequence was then used to as the seed sequence to identify putatively homologous sequences in the archaeabacterial genomes following the selection procedure outlined above. The procedure was repeated for the eubacterial genomes with the selected archaeabacterial sequences as the seed sequences.

### P4 analyses of individual proteins

MCMC analyses of standard amino-acid coded data were conducted with 120,000 generations after the initial burn-in period, sampling trees and parameters every 120 generations. Model parameter proposal tuning values were determined using two independent chains which were each tuned after intervals of 5,000 generations using the P4 "autoTune" method until the tuning values converged between the chains. The burn-in was determined by calculating the cumulative average standard deviation of splits and comparing split support in consensus trees between the two independent chains at intervals of 2,000 generations. When the value of the cumulative average standard deviation of splits was less than 0.08 and when all splits  $\geq 95\%$  in the consensus tree of one chain were supported by  $\geq 70\%$  in the consensus tree of the second chain, the chains were assumed to have reached stationarity. After chain completion, plots of the log likelihood values against generation were checked to ensure that the likelihoods had plateaued. Model adequacy with respect to composition was tested using posterior predictive simulations of the  $\chi^2$  homogeneity statistic at each of 10,000 sample points.

### Bayes Factor calculations

The Bayes factor is the ratio of the marginal likelihoods of the two models being compared, which we use as  $2\log_e(BF)$ , twice the difference in the logs of the marginal likelihoods. These are often interpreted using rough tables such as that given in Kass and Raftery (1), which says that a  $2\log_e(BF)$  of greater than 10 is considered very strong evidence in favor of the better model. While the marginal likelihoods are commonly estimated using the harmonic mean estimator, it may occasionally be unstable (1, 2). For this reason we used the estimator described in equation 16 in Newton and Raftery (2).

### Note on Figures

Individual gene analyses in Figures S1-S51 appear in the same order as the genes are listed in Table S2. Articles cited in the table and figure legends are detailed below.

1. Kass RE, Raftery AE (1995) Bayes Factors. *J Amer Stat Assoc* 90:773–795.
2. Newton MA, Raftery AE (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society* 56:3–56.

3. Matsuzaki M, et al. (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653–657.
4. Bult CJ, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science* 273:1058–1073.
5. Whelan S, Goldman N (2001) A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach *Mol Biol Evol* 18:691–699.
6. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins, vol. 5. National Biomedical Research Foundation, Washington D.C.
7. Embley TM, Martin W (2006) Eukaryotic evolution, changes and challenges. *Nature* 440:623–630.

Taxonomy	Taxon	Provenance
<b>Eubacteria</b>		
Aquificales	<i>Aquiflex aeolicus</i>	GB: NC_000918
	<i>Campylobacter jejuni</i>	GB: NC_003912
	<i>Chlamydia trachomatis</i>	GB: NC_000117
	<i>Clostridium acetobutylicum</i>	GB: NC_003030
	<i>Escherichia coli</i>	GB: NC_000913
	<i>Rickettsia prowazekii</i>	GB: NC_000963
	<i>Rhodopirellula baltica</i>	GB: NC_005027
	<i>Rhodopseudomonas palustris</i>	GB: NC_005296
	<i>Synechocystis sp. PCC6803</i>	GB: NC_000911
	<i>Treponema pallidum</i>	GB: NC_000919
<b>Archaeabacteria</b>		
Crenarchaeota/eocyte	<i>Aeropyrum pernix</i>	GB: NC_000854
	<i>Archaeoglobus fulgidus</i>	GB: NC_000917
	<i>Halobacterium salinarum</i>	GB: NC_002607
	<i>Methanococcus jannaschii</i>	GB: NC_000909
	<i>Methanopyrus kandleri</i>	GB: NC_003551
	<i>Methanoscincina mazei</i>	GB: NC_003901
	<i>Methanothermobacter thermautotrophicus</i>	GB: NC_000916
	<i>Nanoarchaeum equitans</i>	GB: NC_005213
	<i>Picrophilus torridus</i>	GB: NC_005877
	<i>Pyrobaculum aerophilum</i>	GB: NC_003364
	<i>Pyrococcus furiosus</i>	GB: NC_003413
	<i>Sulfolobus solfataricus</i>	GB: NC_002754
	<i>Thermococcus kodakaraensis</i>	GB: NC_006624
	<i>Thermoplasma acidophilum</i>	GB: NC_002578
<b>Eukaryotes</b>		
Viridiplantae	<i>Arabidopsis thaliana</i>	TIGR <sup>a</sup>
	<i>Cryptococcus neoformans</i>	STGC <sup>b</sup>
	<i>Cryptosporidium hominis</i>	VCU <sup>c</sup>
	<i>Cyanidioschyzon merolae</i>	C. m. Genome Project <sup>d</sup>
	<i>Dictyostelium discoideum</i>	dictyBase <sup>e</sup>
	<i>Drosophila melanogaster</i>	EMBL-EBI <sup>f</sup>
	<i>Encephalitozoon cuniculi</i>	GB: nr <sup>g</sup>
	<i>Entamoeba histolytica</i>	TIGR <sup>a</sup>
	<i>Giardia lamblia</i>	GiardiaDB <sup>h</sup>
	<i>Homo sapiens</i>	EMBL-EBI <sup>f</sup>
	<i>Leishmania major</i>	WTSI <sup>i</sup>
	<i>Phytophthora ramorum</i>	JGI <sup>j</sup>
	<i>Plasmodium falciparum</i>	TIGR <sup>a</sup>
	<i>Saccharomyces cerevisiae</i>	GB: nr <sup>g</sup>
	<i>Trichomonas vaginalis</i>	TIGR <sup>a</sup>
	<i>Thalassiosira pseudonana</i>	JGI <sup>j</sup>

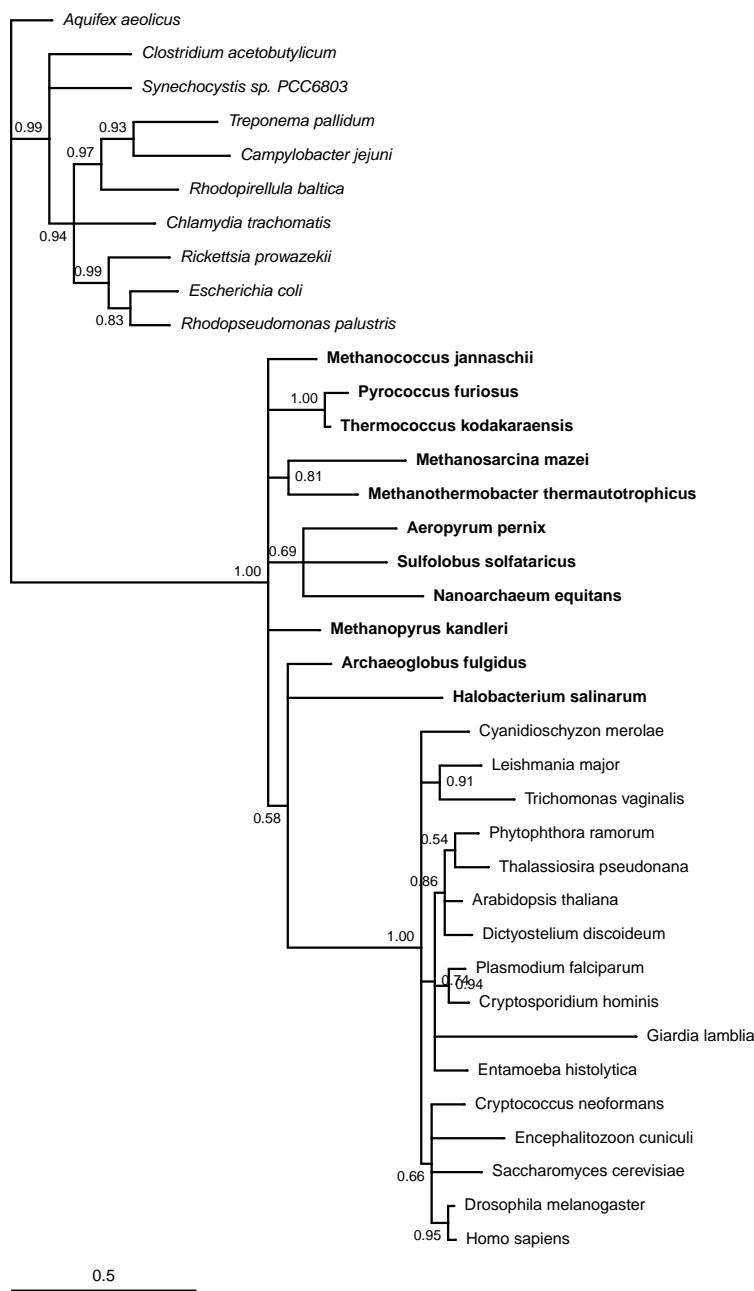
**Table S1:** Taxa and data provenance. <sup>a</sup>The Institute for Genomic Research, <sup>b</sup>Standford Genome Technology Centre, <sup>c</sup>Virginia Commonwealth University, Center for the Study of Biological Complexity., <sup>d</sup><http://merolae.biol.s.u-tokyo.ac.jp/> (3), <sup>e</sup><http://www.dictybase.org/>, <sup>f</sup>European Molecular Biology Laboratory - European Bioinformatics Institute, <sup>g</sup><ftp://ftp.ncbi.nih.gov/blast/> - NCBI non-redundant protein BLAST database, <sup>h</sup><http://gmod.mbl.edu/perl/site/giardia14>, <sup>i</sup>Wellcome Trust Sanger Institute, <sup>j</sup>Joint Genome Institute.

F.c. <sup>a</sup>	Locus	nTax <sup>b</sup>	nChar <sup>c</sup>	Model <sup>d</sup>
I	26S proteasome subunit 1	37	231	WAG+I+Γ+2CV
I	*26S proteasome subunit 8	36	236	WAG+I+Γ+2CV
I	40S ribosomal protein S18 (S13)	40	97	WAG+I+Γ+3CV
I	40S ribosomal protein S11 (S17)	39	60	Dayhoff+I+Γ+1CV
I	40S ribosomal protein S14 (S11)	39	118	WAG+Γ+1CV
I	40S ribosomal protein S16 (S9)	39	82	WAG+Γ+2CV
I	40S ribosomal protein S2 (S5)	40	106	WAG+Γ+2CV
I	40S ribosomal protein S23 (S12)	39	69	WAG+I+Γ+1CV
I	40S ribosomal protein S5 (S7)	40	112	WAG+I+Γ+2CV
I	40S ribosomal protein S3	40	116	WAG+I+Γ+4CV
I	40S ribosomal protein SA (P40,S2)	39	127	WAG+I+Γ+1CV
I	60S ribosomal protein L10 (L10e,L16)	40	64	WAG+I+Γ+2CV
I	60S ribosomal protein L10A (L1)	40	97	WAG+Γ+4CV
I	60S ribosomal protein L11 (L5)	40	109	WAG+Γ+1CV
I	60S ribosomal protein L17 (L22)	40	63	WAG+I+Γ+4CV
I	60S ribosomal protein L12 (L11)	39	99	WAG+Γ+3CV
I	60S ribosomal protein L23 (L14)	40	90	WAG+I+Γ+2CV
I	60S ribosomal protein L8 (L2)	40	159	WAG+I+Γ+2CV
I	60S ribosomal protein L3	40	63	WAG+I+Γ+1CV
I	60S ribosomal protein L13A	40	65	WAG+I+Γ+1CV
I	60S ribosomal protein L5 (L18)	40	64	WAG+I+Γ+3CV
I	60S ribosomal protein L4	39	68	WAG+I+Γ+3CV
O	ATP-binding cassette E-1	38	134	WAG+I+Γ+2CV
I	Elongation Factor 2 (EF-G)	40	304	WAG+I+Γ+3CV
I	Elongation Factor 1α (EF Tu)	38	202	WAG+I+Γ+2CV
I	RNA polymerase I RPA1	38	332	WAG+I+Γ+2CV
I	RNA polymerase I RPA2	40	222	WAG+I+Γ+1CV
I	*RNA polymerase II RPB1	39	432	WAG+I+Γ+2CV
I	*RNA polymerase II RPB2	40	313	WAG+I+Γ+1CV
I	*RNA polymerase III RPC1	39	377	WAG+I+Γ+2CV
I	*RNA polymerase III RPC2	39	267	WAG+I+Γ+2CV
O	V-type ATPase V1 subunit B	38	285	WAG+I+Γ+4CV
O	Chaperonin containing TCP1 1 (α)	40	202	WAG+I+Γ+2CV
O	Chaperonin containing TCP1 3 (γ)	40	175	WAG+I+Γ+2CV
O	Chaperonin containing TCP1 4 (δ)	37	218	WAG+I+Γ+2CV
O	Chaperonin containing TCP1 5 (ε)	40	187	WAG+Γ+2CV
O	Chaperonin containing TCP1 7 (η)	40	165	WAG+I+Γ+2CV
I	Glutamine-tRNA ligase	39	98	WAG+I+Γ+1CV
I	*Glutamate-tRNA ligase	38	137	WAG+I+Γ+2CV
I	Aspartate-tRNA ligase	40	223	WAG+I+Γ+4CV
I	Methionyl aminopeptidase	39	111	WAG+I+Γ+4CV
O	Protein transport protein Sec61α	40	102	WAG+I+Γ+1CV
O	Transitional endoplasmic reticulum ATPase	40	213	WAG+I+Γ+2CV
O	Pseudouridine syntase component dyskerin	39	111	WAG+I+Γ+2CV
I	Phenylalanine-tRNA ligase (β)	39	97	WAG+I+Γ+9CV
I	O-sialoglycoprotein endopeptidase	39	176	WAG+I+Γ+6CV
I	Translation initiation factor IF-2	40	187	WAG+I+Γ+4CV
I	Replication factor C subunit 2	39	141	WAG+I+Γ+2CV
I	Replication factor C subunit 4	39	119	WAG+I+Γ+2CV
O	Signal recognition particle SR-α	39	159	WAG+I+Γ+4CV
O	Signal recognition particle SRP54	39	243	WAG+I+Γ+9CV <sup>a</sup>
I	*Ribosomal RNA large subunit	35	686 <sup>e</sup>	GTR+I+Γ+2CV
I	*Ribosomal RNA small subunit	39	362 <sup>e</sup>	GTR+I+Γ+2CV

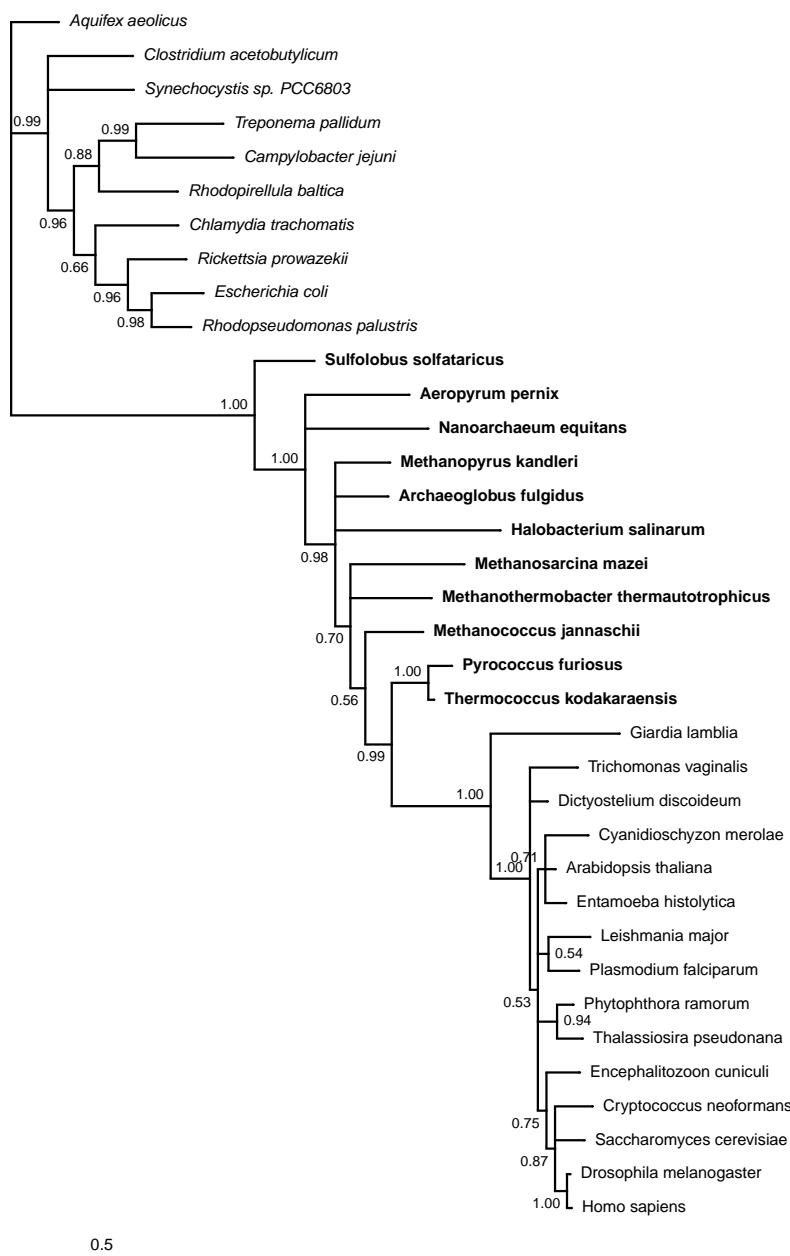
**Table S2:** Gene loci included in the analyses, data statistics, and optimal composition model for the locus. <sup>a</sup>functional class of protein: I, informational, and O, operational (4); <sup>b</sup>number of taxa, <sup>c</sup>number of characters included, <sup>d</sup> optimal model, <sup>e</sup>included maximum likelihood informative characters only. Model abbreviations: WAG, substitution rate model of Whelan and Goldman (5); Dayhoff, substitution rate model of Dayhoff (6); I, proportion of invariant sites; Γ, gamma-distributed among site rate variation (4 categories); CV: optimum number of composition vectors. Loci prepended by an asterisk were not included in the 45 locus combined analyses. <sup>a</sup>9 composition vectors was not optimal P=0.0342.

Locus	$\log_e(L_m)$ hom <sup>a</sup>	$\log_e(L_m)$ opt <sup>b</sup>	$\log_e(L_m)$ cov <sup>c</sup>	$z\log_e(B_{hom cov})^d$	$z\log_e(B_{opt cov})^e$
26S proteasome subunit 1	-7272.39	-7255.86	-7312.65	80.53	<b>113.59</b>
26S proteasome subunit 8	-7063.42	-7038.39	-7105.48	84.12	<b>134.19</b>
40S ribosomal protein S18 (S13)	-5099.06	-5048.98	-5125.91	53.71	<b>153.87</b>
40S ribosomal protein S11 (S17)	-2970.25	-2988.08	-3060.16	179.81	<b>144.17</b>
40S ribosomal protein S14 (S11)	-4230.12	-4245.94	-4234.07	7.89	-23.76
40S ribosomal protein S16 (S9)	-3790.55	-3779.61	-3843.49	105.88	<b>127.75</b>
40S ribosomal protein S2 (S5)	-4933.58	-4904.16	-4934.63	2.11	<b>60.93</b>
40S ribosomal protein S23 (S12)	-1931.72	-1942.55	-1962.44	61.45	<b>39.77</b>
40S ribosomal protein S5 (S7)	-5327.67	-5314.57	-5303.30	<b>-48.74</b>	-22.54
40S ribosomal protein S3	-6222.72	-6174.24	-6287.58	129.73	<b>226.68</b>
40S ribosomal protein SA (P40,S2)	-6312.26	-6326.74	-6343.42	62.32	<b>33.37</b>
60S ribosomal protein L10 (L10e,L16)	-2966.16	-2952.14	-3003.38	74.43	<b>102.47</b>
60S ribosomal protein L10A (L1)	-5539.64	-5480.25	-5537.30	<b>-4.68</b>	<b>114.09</b>
60S ribosomal protein L11 (L5)	-5094.97	-5110.31	-5093.25	<b>-3.44</b>	-34.11
60S ribosomal protein L17 (L22)	-3876.46	-3839.05	-3917.50	82.09	<b>156.90</b>
60S ribosomal protein L12 (L11)	-4812.41	-4778.50	-4832.81	40.79	<b>108.61</b>
60S ribosomal protein L23 (L14)	-3537.57	-3534.41	-3582.49	89.84	<b>96.16</b>
60S ribosomal protein L8 (L2)	-7745.27	-7728.14	-7728.99	<b>-32.55</b>	<b>1.70</b>
60S ribosomal protein L3	-2706.41	-2716.32	-2749.00	85.18	<b>65.36</b>
60S ribosomal protein L13A	-3226.82	-3237.73	-3265.47	77.29	<b>55.46</b>
60S ribosomal protein L5 (L18)	-3244.81	-3212.55	-3261.74	33.87	<b>98.38</b>
60S ribosomal protein L4	-3845.31	-3801.52	-3885.36	80.10	<b>167.68</b>
ATP-binding cassette E-1	-4952.46	-4927.07	-4980.00	55.08	<b>105.87</b>
Elongation Factor 2 (EF-G)	-9689.69	-9616.60	-9686.09	<b>-7.20</b>	<b>139.00</b>
Elongation Factor 1 $\alpha$ (EF Tu)	-5651.79	-5634.75	-5675.23	46.88	<b>80.95</b>
RNA polymerase I RPA1	-12793.85	-12747.08	-12795.00	2.31	<b>95.85</b>
RNA polymerase I RPA2	-8146.17	-8166.91	-8183.48	74.63	<b>33.14</b>
RNA polymerase II RPB1	-16405.93	-16351.20	-16402.59	<b>-6.68</b>	<b>102.78</b>
RNA polymerase II RPB2	-11415.68	-11428.97	-11553.61	275.86	<b>249.27</b>
RNA polymerase III RPC1	-13548.38	-13495.33	-13516.68	<b>-63.40</b>	<b>42.69</b>
RNA polymerase III RPC2	-8828.34	-8805.17	-8875.20	93.72	<b>140.06</b>
V-type ATPase V1 subunit B	-10419.16	-10302.89	-10371.26	<b>-95.79</b>	<b>136.75</b>
Chaperonin containing TCP1 1 ( $\alpha$ )	-8672.42	-8622.74	-8666.85	<b>-11.15</b>	<b>88.21</b>
Chaperonin containing TCP1 3 ( $\gamma$ )	-7829.63	-7794.34	-7835.94	12.62	<b>83.19</b>
Chaperonin containing TCP1 4 ( $\delta$ )	-9428.54	-9390.09	-9502.18	147.27	<b>224.19</b>
Chaperonin containing TCP1 5 ( $\epsilon$ )	-8375.95	-8336.98	-8368.15	<b>-15.59</b>	<b>62.34</b>
Chaperonin containing TCP1 7 ( $\eta$ )	-7234.67	-7205.62	-7236.31	3.28	<b>61.38</b>
Glutamine-tRNA ligase	-4565.85	-4576.04	-4596.28	60.86	<b>40.48</b>
Glutamate-tRNA ligase	-6968.82	-6917.91	-7038.54	139.46	<b>241.26</b>
Aspartate-tRNA ligase	-11182.42	-11108.43	-11220.25	75.66	<b>223.63</b>
Methionyl aminopeptidase	-5782.04	-5726.24	-5824.55	85.03	<b>196.63</b>
Protein transport protein Sec61 $\alpha$	-4337.56	-4351.18	-4439.52	203.93	<b>176.69</b>
Transitional endoplasmic reticulum ATPase	-6116.70	-6086.57	-6116.20	<b>-1.02</b>	<b>59.25</b>
Pseudouridine syntase component dyskerin	-4595.70	-4562.85	-4608.35	25.29	<b>91.00</b>
Phenylalanine-tRNA ligase ( $\beta$ )	-5708.62	-5608.09	-5752.71	88.18	<b>289.23</b>
O-sialoglycoprotein endopeptidase	-8159.28	-8046.89	-8187.25	55.96	<b>280.73</b>
Translation initiation factor IF-2	-7775.11	-7702.81	-7832.19	114.15	<b>258.75</b>
Replication factor C subunit 2	-6694.35	-6673.16	-6757.06	125.42	<b>167.80</b>
Replication factor C subunit 4	-5746.41	-5712.08	-5778.90	64.97	<b>133.63</b>
Signal recognition particle SR- $\alpha$	-7646.35	-7575.37	-7699.53	106.37	<b>248.32</b>
Signal recognition particle SRP54	-12267.43	-12111.48	-12356.81	178.76	<b>490.67</b>
Combined SSU and LSU rRNA	-23960.00	-23507.36	-23637.60	<b>-644.80</b>	<b>260.48</b>

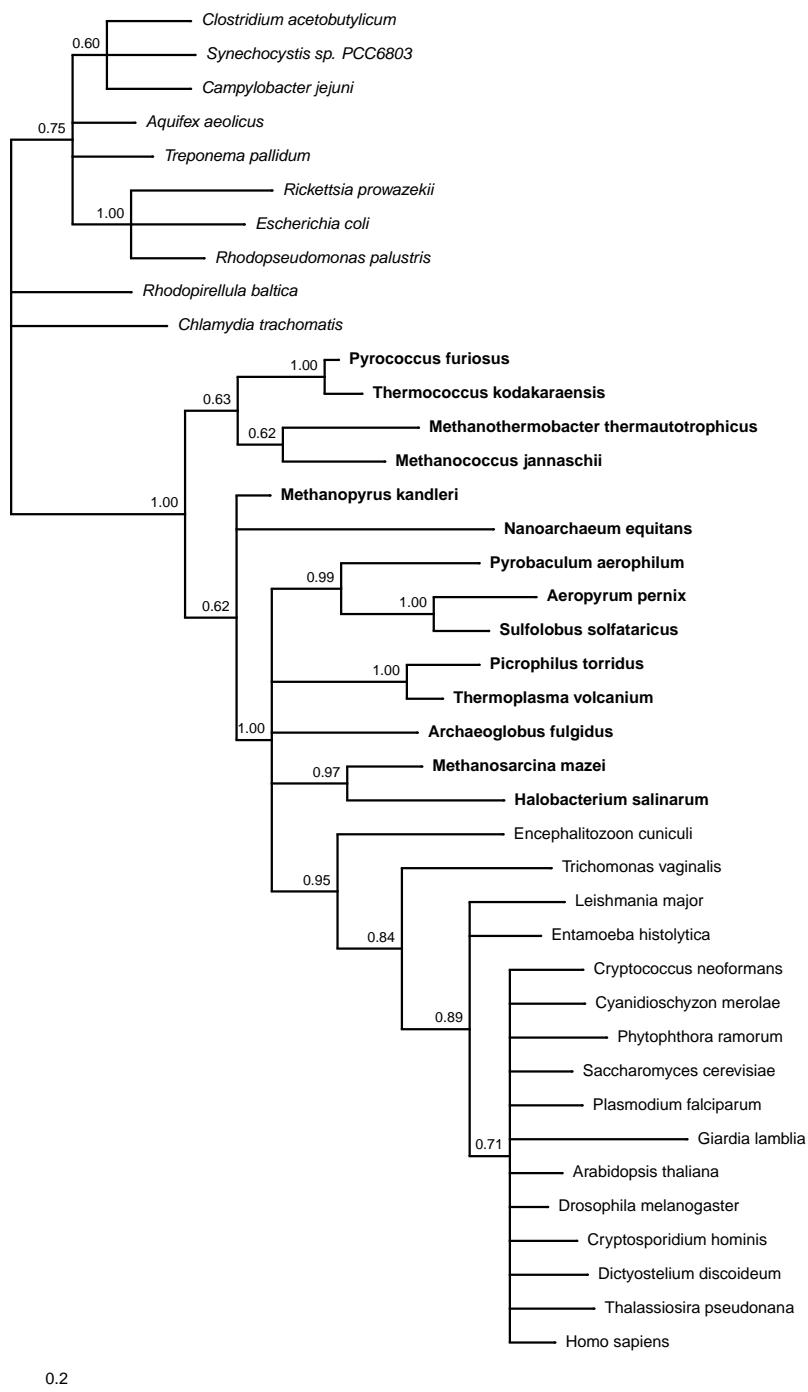
**Table S3:** Logs of estimates of marginal likelihoods (eqn. 16 in ref. (2)) from the posterior distributions and twice log Bayes factor comparisons among models for individual locus analyses. Optimal rate matrix models are provided in Table S2. <sup>a</sup>The log marginal likelihoods of the homogeneous analyses (without the polytomy prior), <sup>b</sup>the log marginal likelihoods of the optimal tree-heterogeneous analyses with the polytomy prior, <sup>c</sup>the log marginal likelihoods of the covarion analyses, <sup>d</sup>twice log Bayes factors comparing the homogeneous analyses versus the covarion, highlighted are those comparisons favouring a covarion over a homogeneous model, and, <sup>e</sup>twice log Bayes factors comparing the optimal tree-heterogeneous analysis with the polytomy prior versus the covarion, highlighted are those comparisons favoring an optimal (and possibly heterogeneous - see Table S2) composition model over a covarion model. Homogeneous analyses were conducted without the polytomy prior so that we were directly able to assess the influence of the covarion (the covarion analyses were conducted in MrBayes, which does not implement the polytomy prior). Higher log likelihoods indicate a better fit to the model. Twice the log Bayes factor of  $>10$  is considered "very strong" evidence in favour of M1 ( $z\log_e(B_{M1|M0}))$ (1)



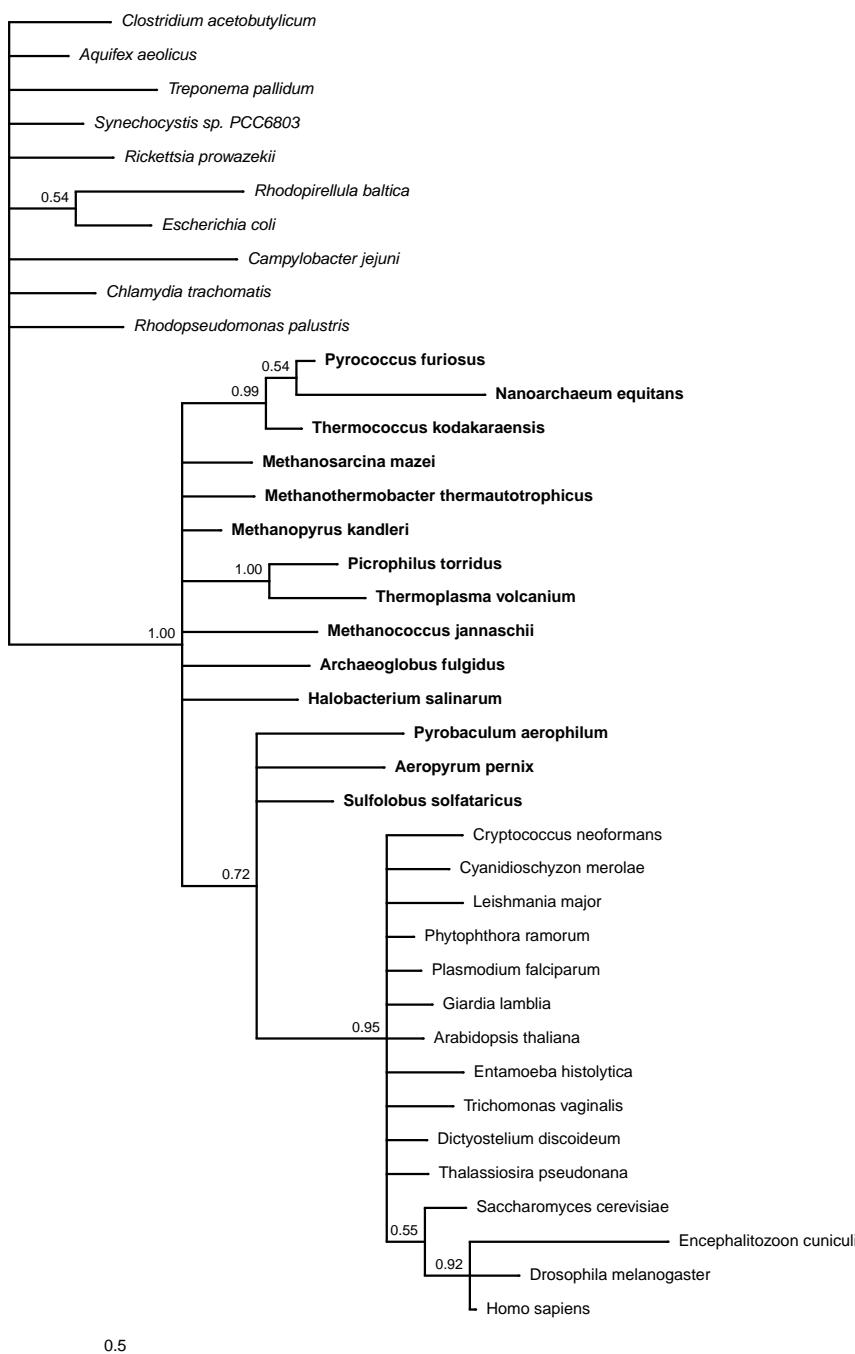
**Fig. S1:** 26S proteasome ATP-dependent regulatory subunit 1 – nTax = 37, nChar = 231 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.3509



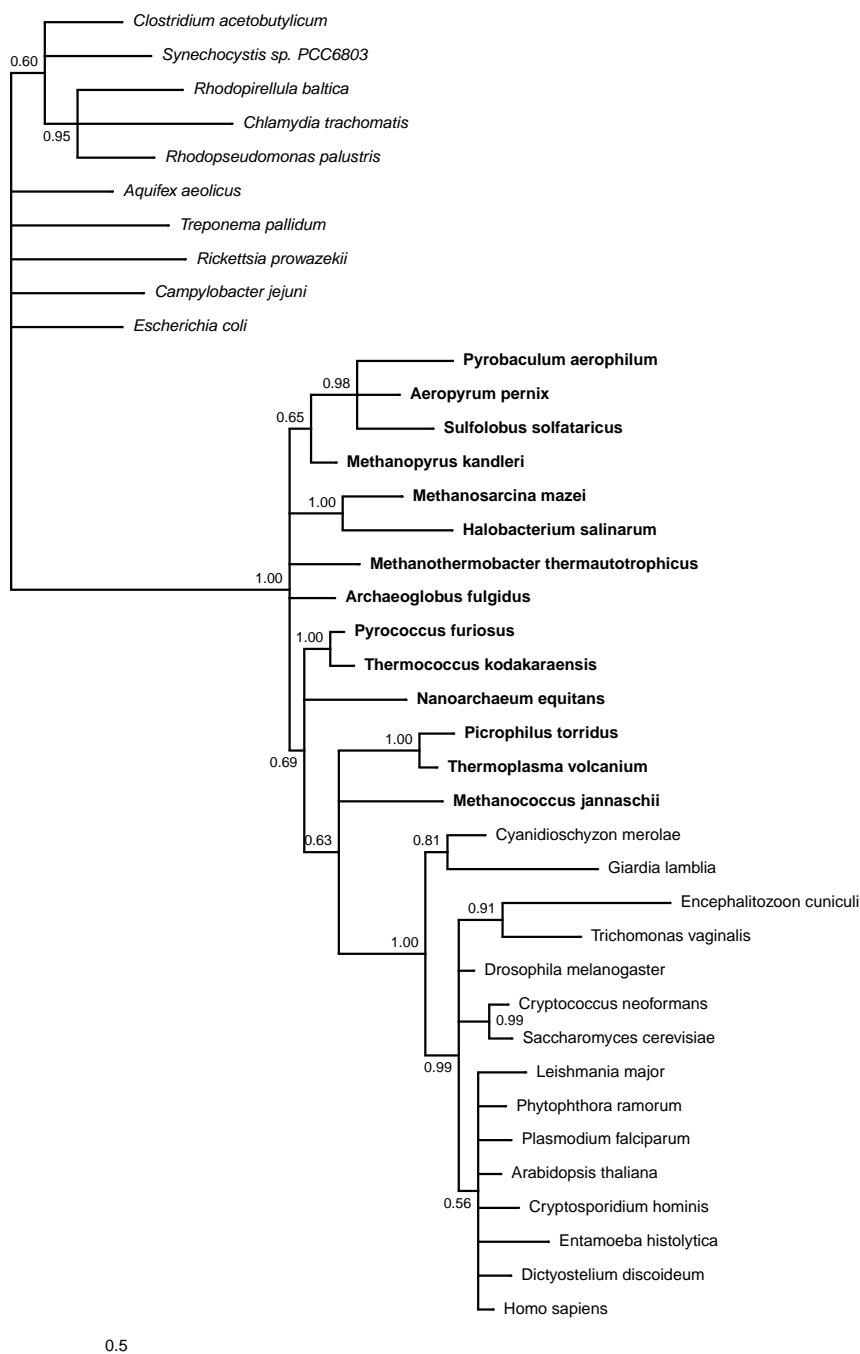
**Fig. S2:** 26S proteasome ATP-dependent regulatory subunit 8 – nTax = 36, nChar = 236 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.0571



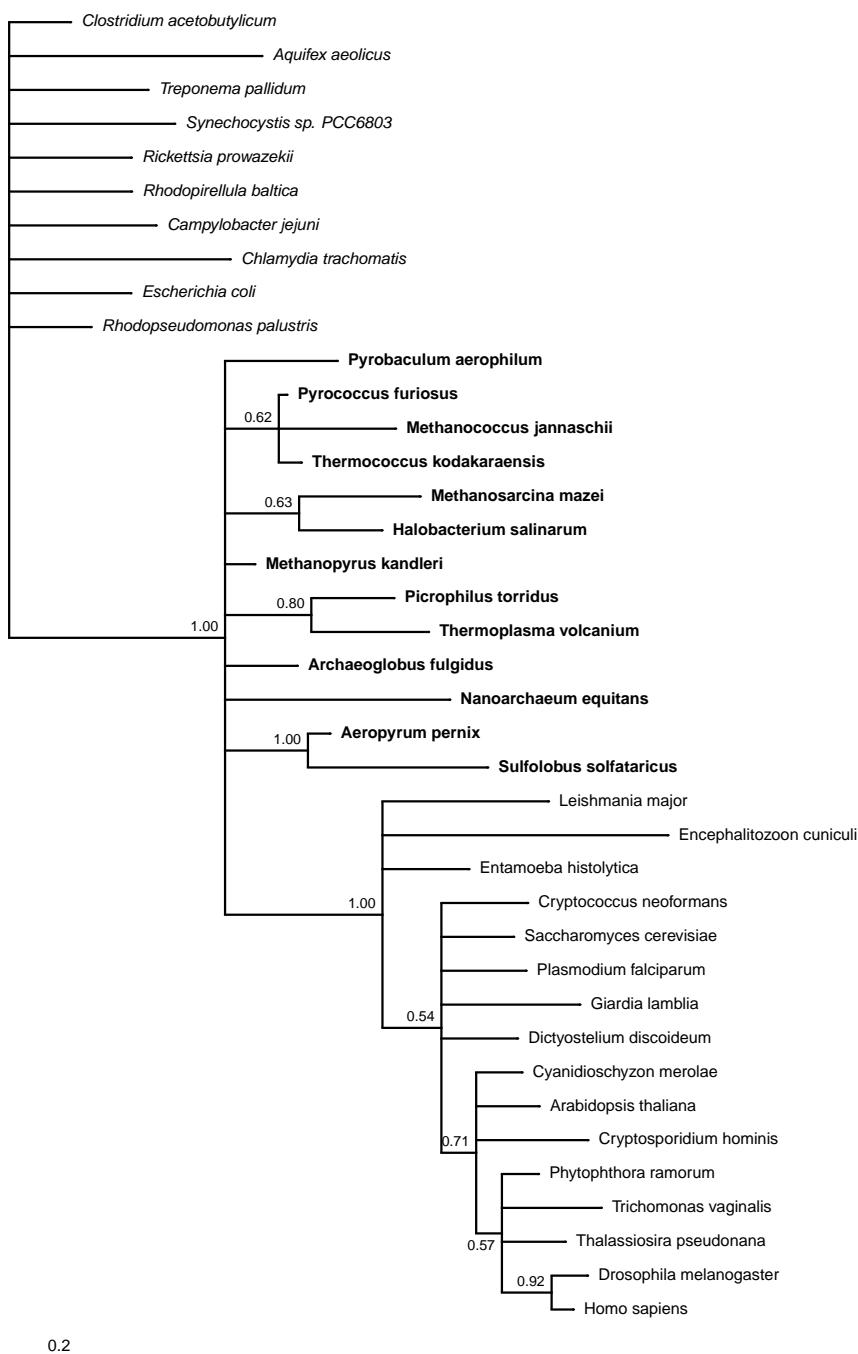
**Fig. S3:** 40S ribosomal protein S18 (S13) – nTax = 40, nChar = 97 Substitution model: WAG+I+Γ+3CV  
Composition homogeneity test P value = 0.0686



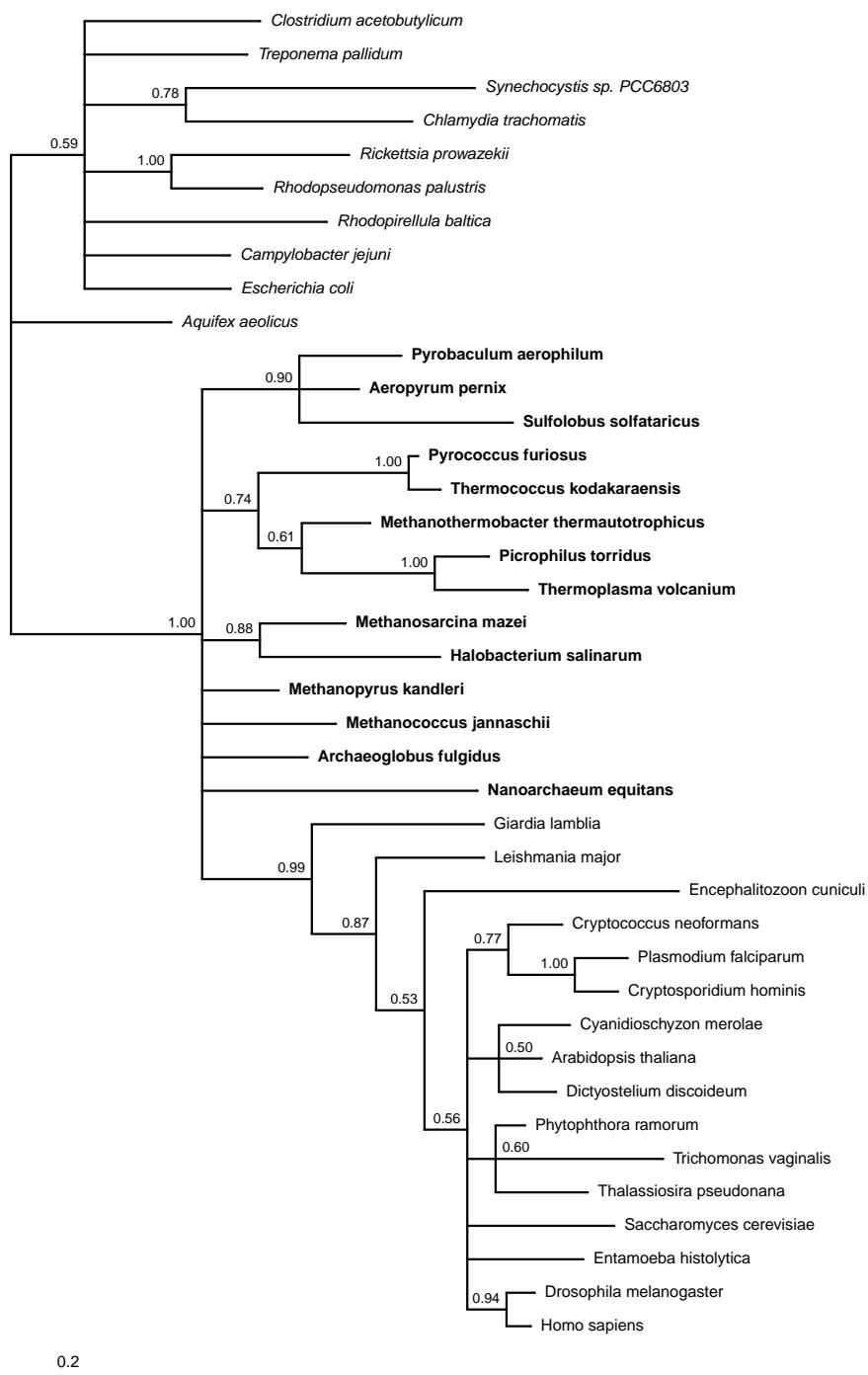
**Fig. S4:** 40S ribosomal protein S11 (S17) – nTax = 39, nChar = 60 Substitution model: Dayhoff+Γ+1CV  
Composition homogeneity test P value = 0.2963



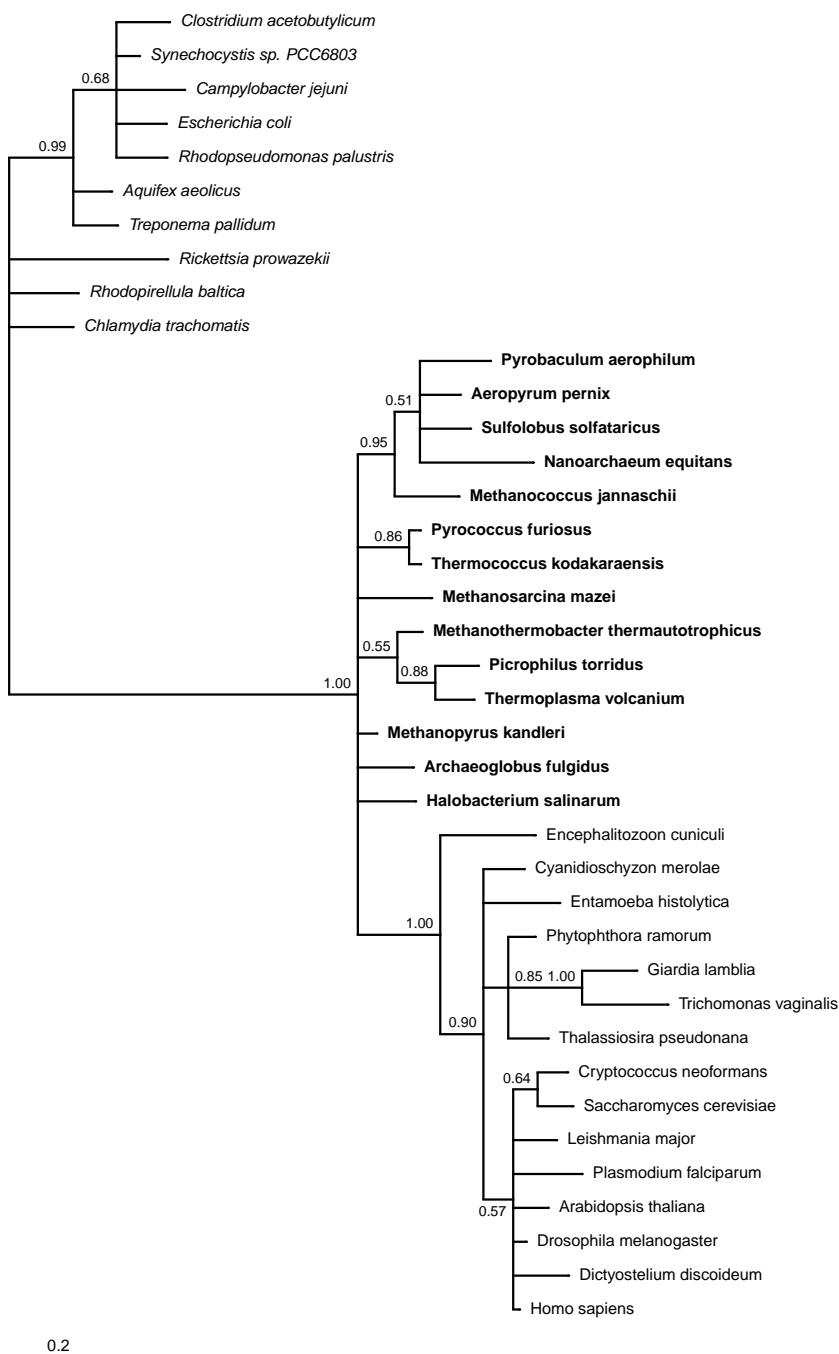
**Fig. S5:** 40S ribosomal protein S14 (S11) – nTax = 39, nChar = 118 Substitution model: WAG+Γ+1CV Composition homogeneity test P value = 0.0852



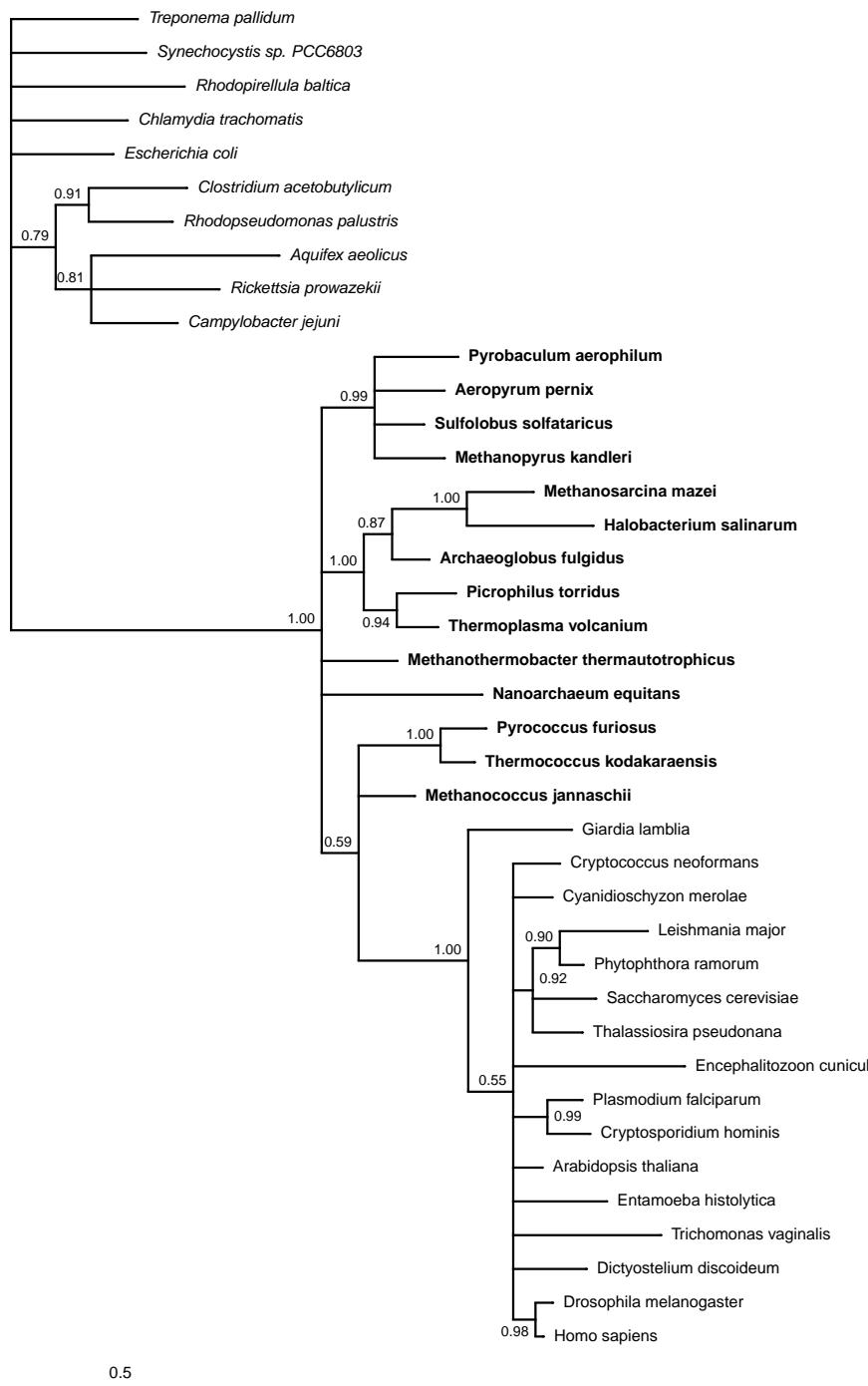
**Fig. S6:** 40S ribosomal protein S16 (S9) – nTax = 39, nChar = 82 Substitution model: WAG+Γ+2CV Composition homogeneity test P value = 0.0833



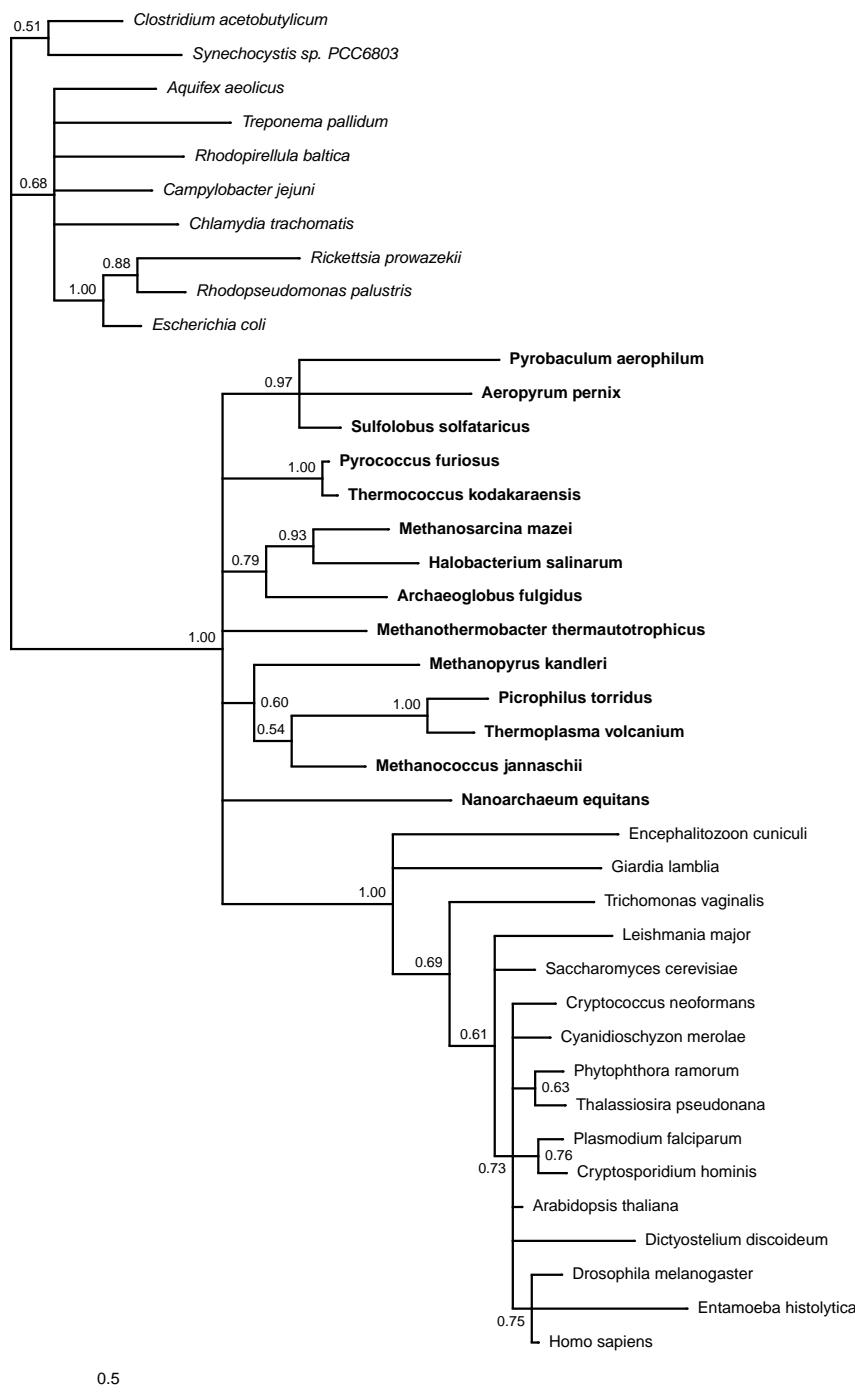
**Fig. S7:** 40S ribosomal protein S2 (S5) – nTax = 40, nChar = 106 Substitution model: WAG+Γ+2CV Composition homogeneity test P value = 0.0539



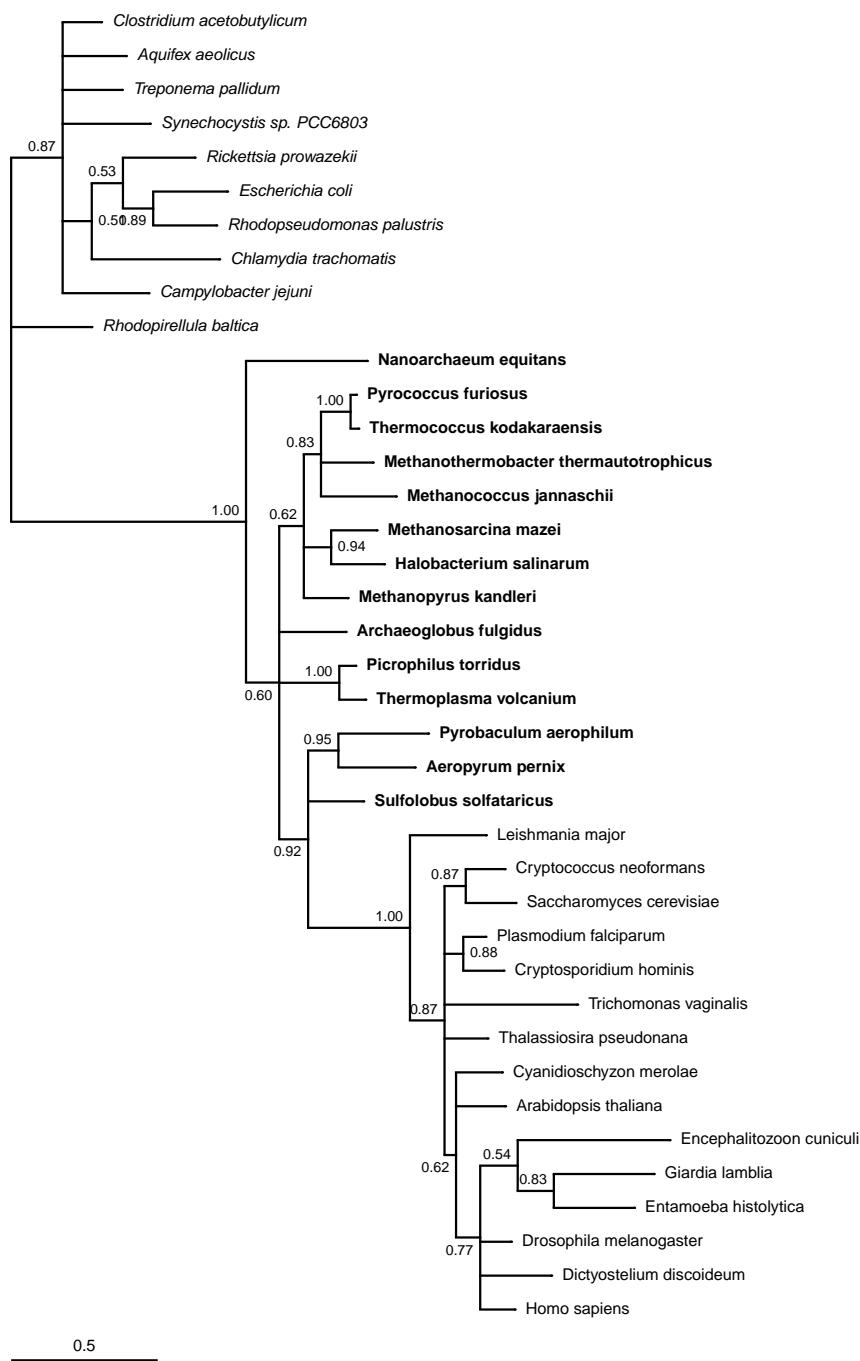
**Fig. S8:** 40S ribosomal protein S23 (S12) – nTax = 39, nChar = 69 Substitution model: WAG+I+Γ+1CV  
Composition homogeneity test P value = 0.2258



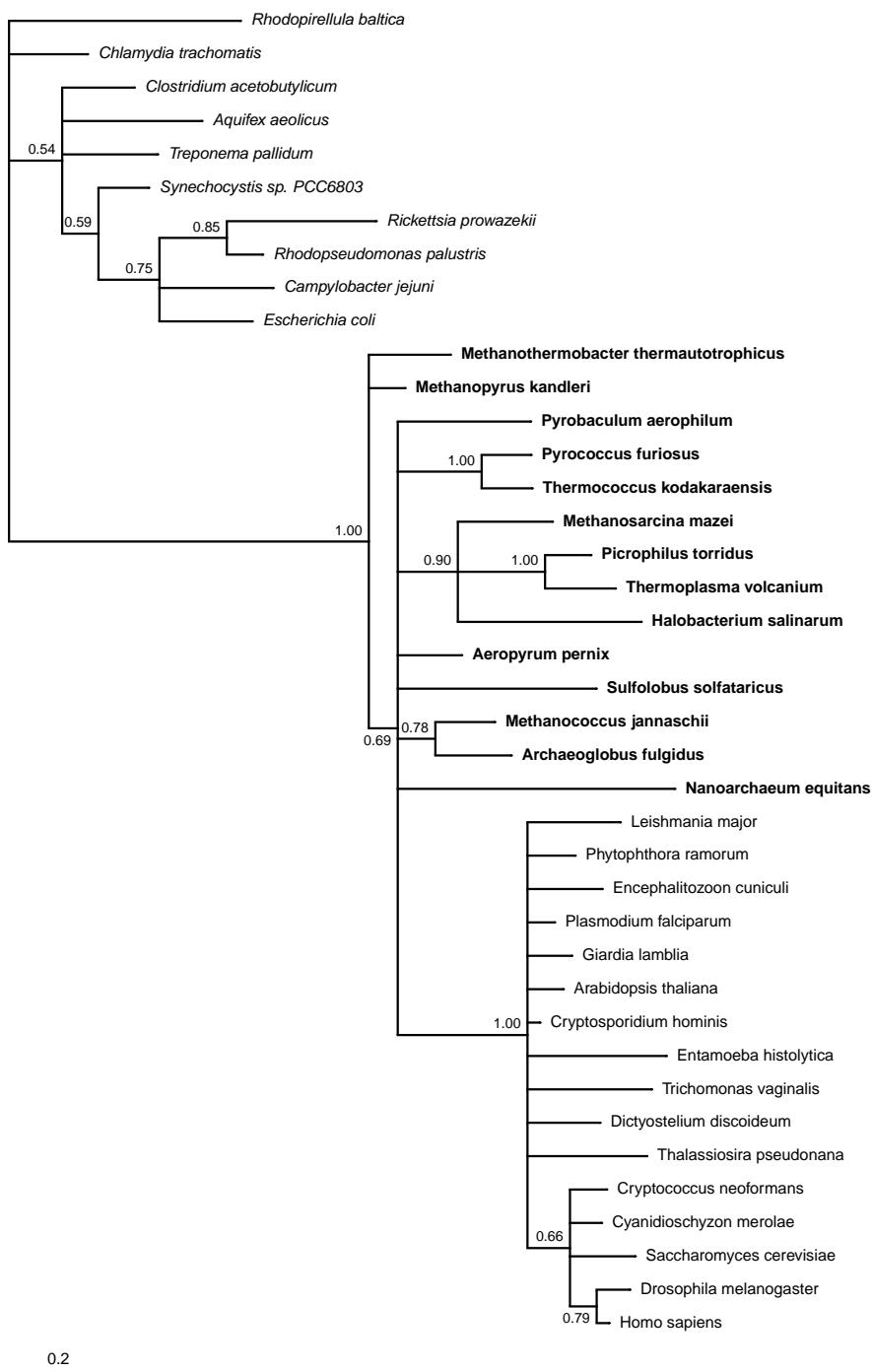
**Fig. S9:** 40S ribosomal protein S5 (S7) – nTax = 40, nChar = 112 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1500



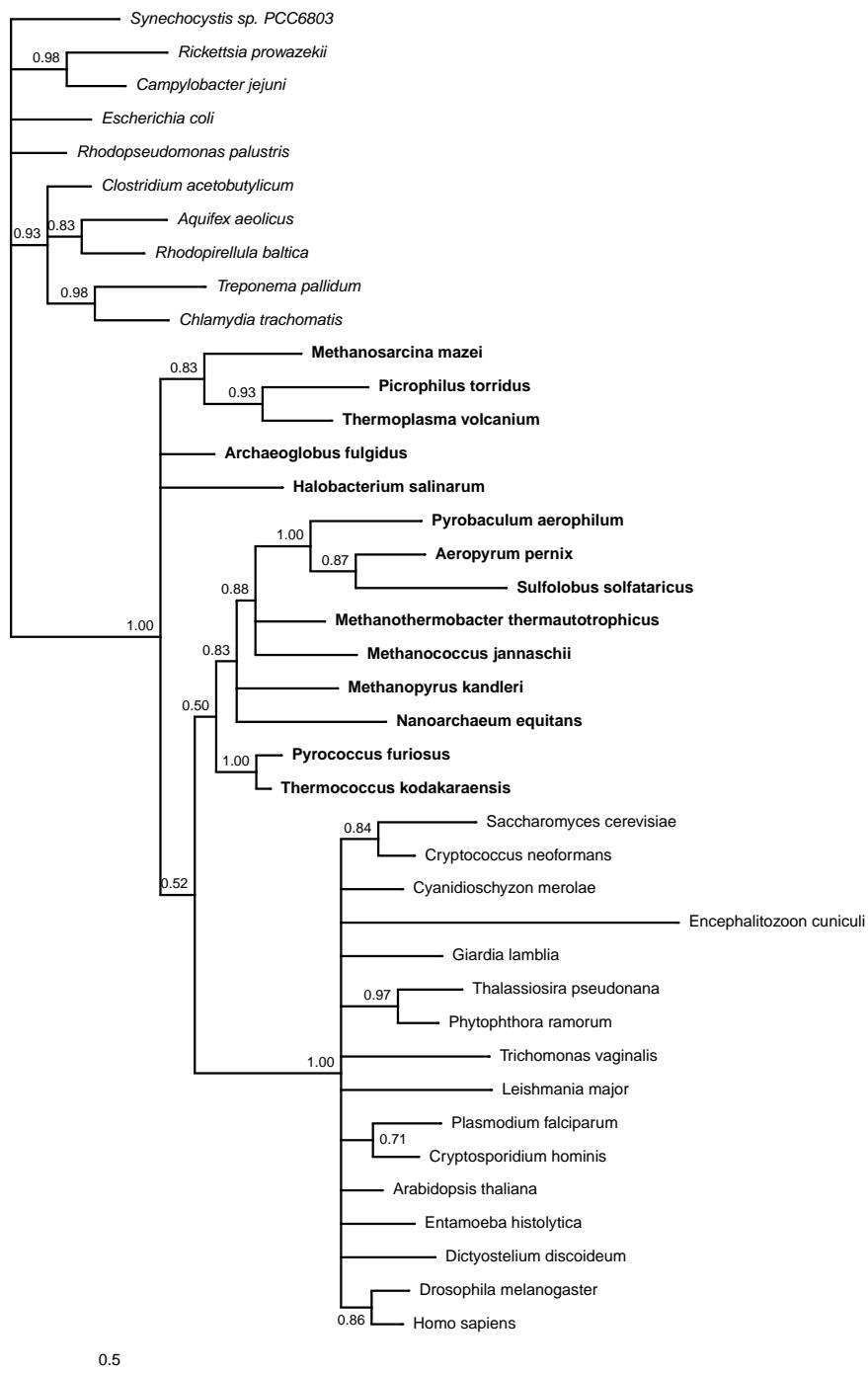
**Fig. S10:** 40S ribosomal protein S<sub>3</sub> – nTax = 40, nChar = 116 Substitution model: WAG+I+Γ+4CV Composition homogeneity test P value = 0.0822



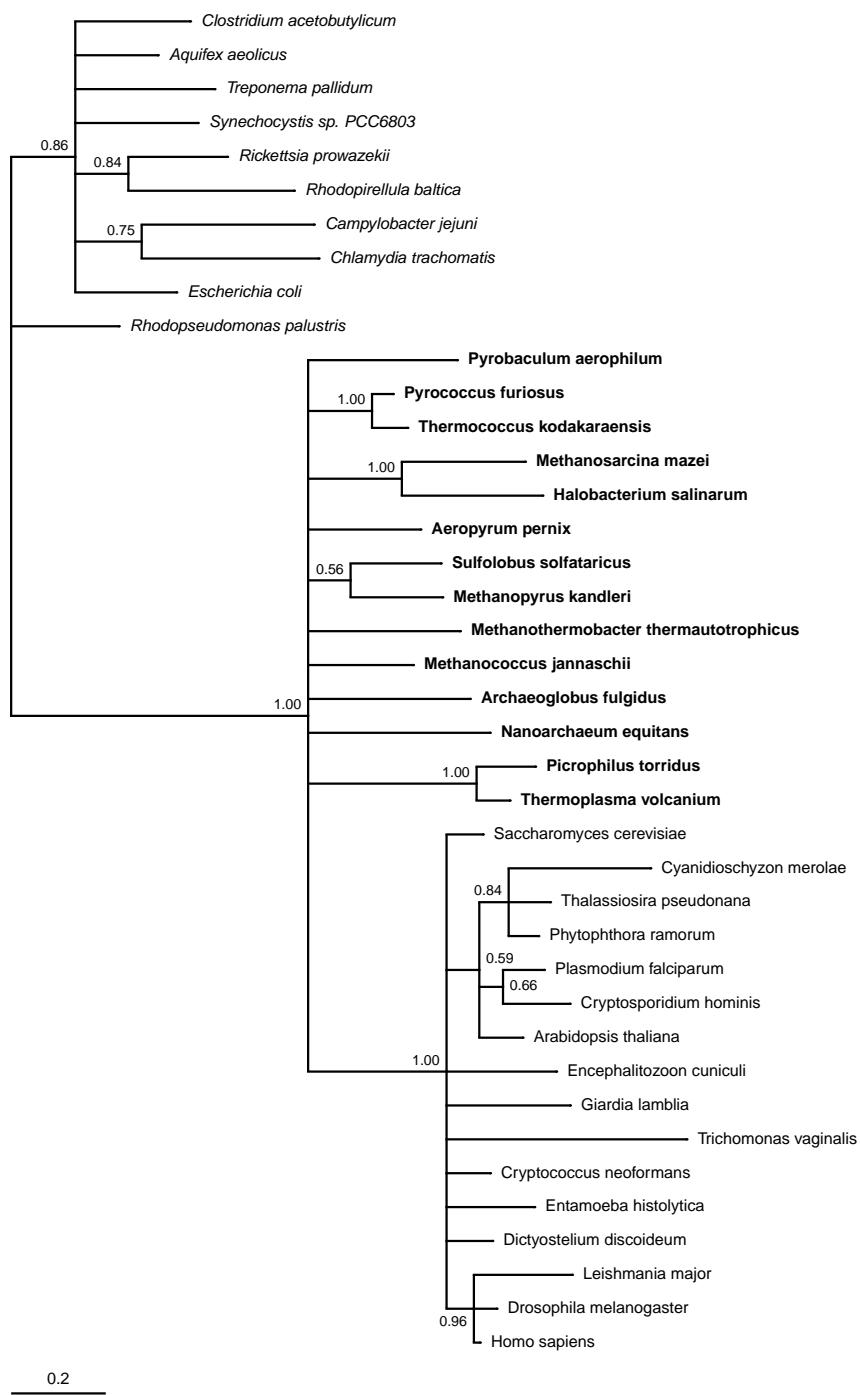
**Fig. S11:** 40S ribosomal protein SA (P40) (S2) – nTax = 39, nChar = 127 Substitution model: WAG+I+Γ+1CV  
Composition homogeneity test P value = 0.0654



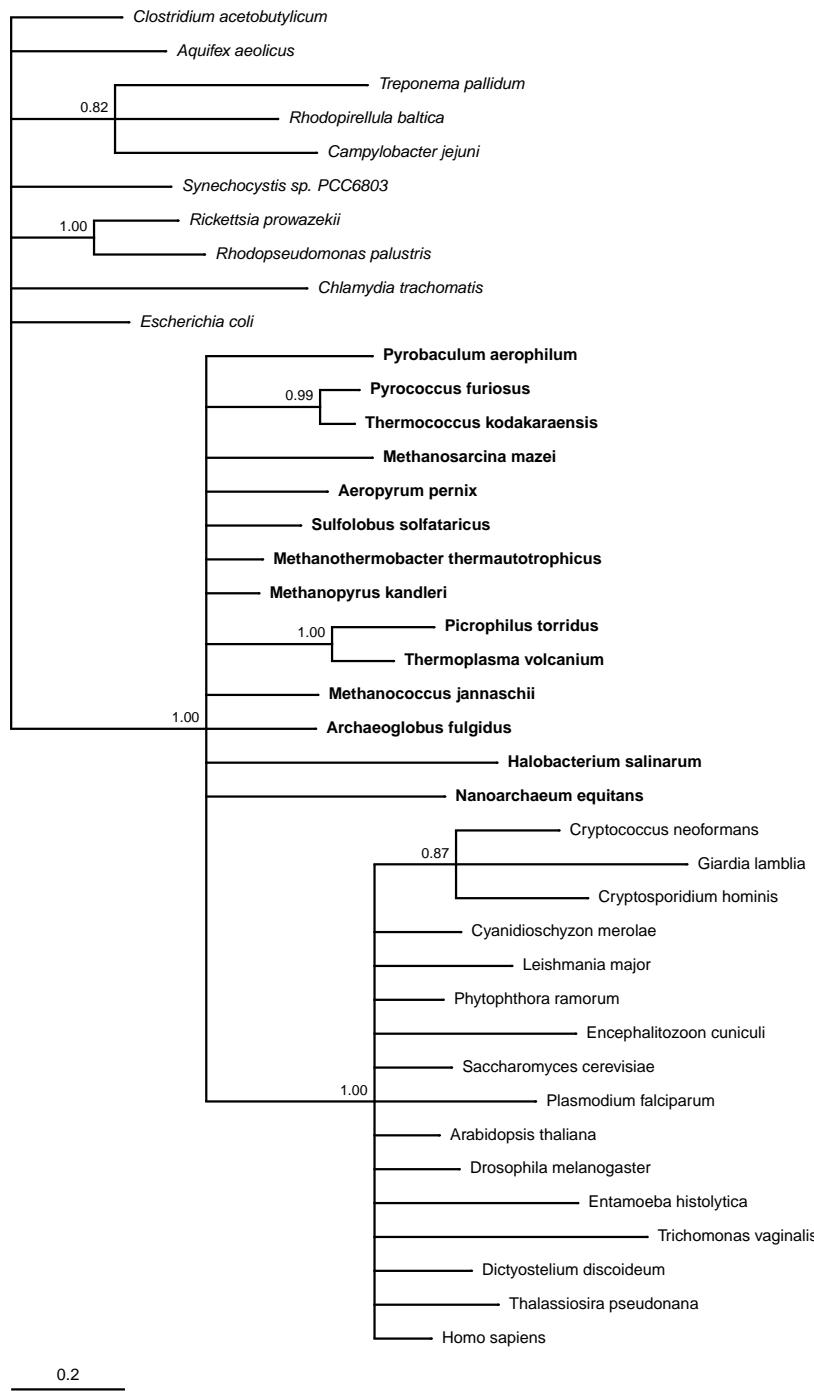
**Fig. S12:** 60S ribosomal protein L10 (L10e, L16) – nTax = 40, nChar = 64 Substitution model: WAG+I+Γ+2CV  
Composition homogeneity test P value = 0.1410



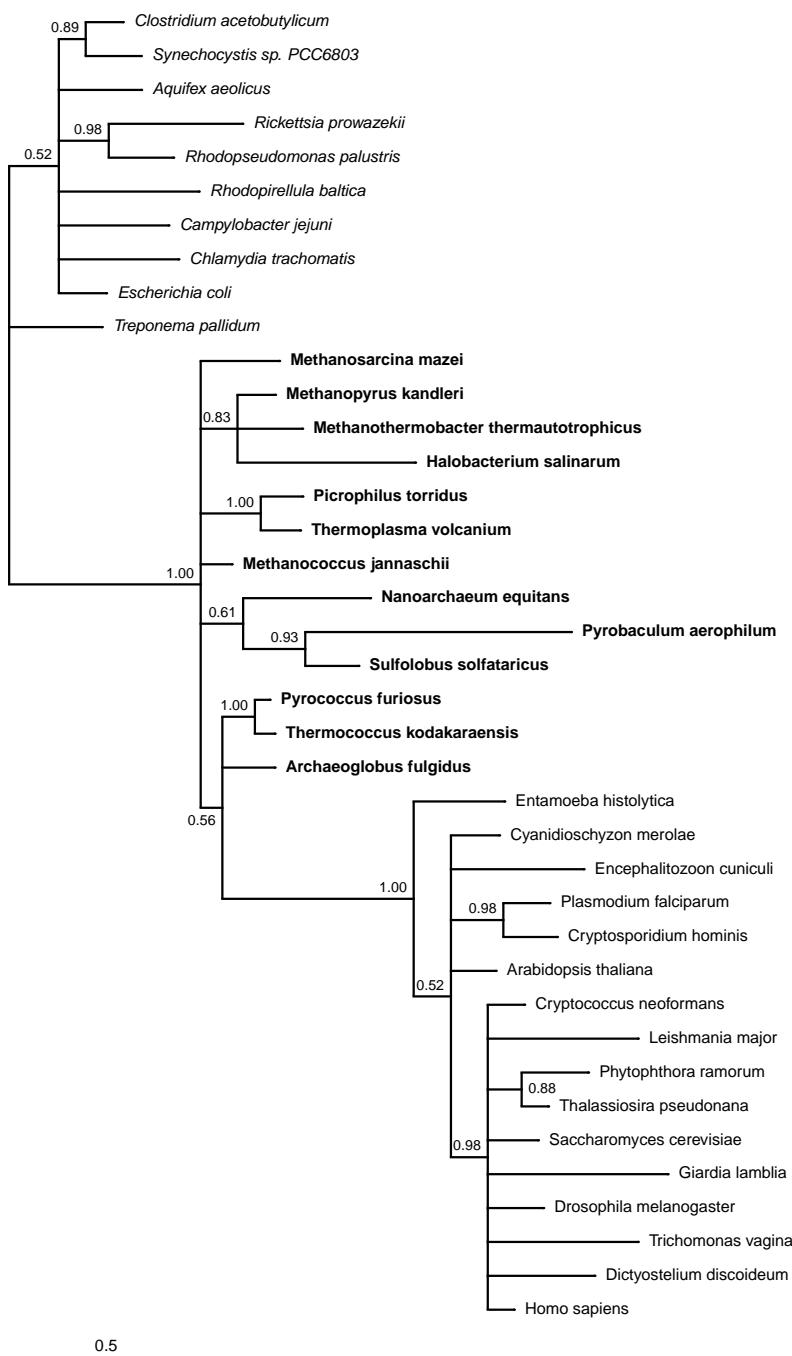
**Fig. S13:** 60S ribosomal protein L10A (L1) – nTax = 40, nChar = 97 Substitution model: WAG+Γ+4CV  
Composition homogeneity test P value = 0.0955



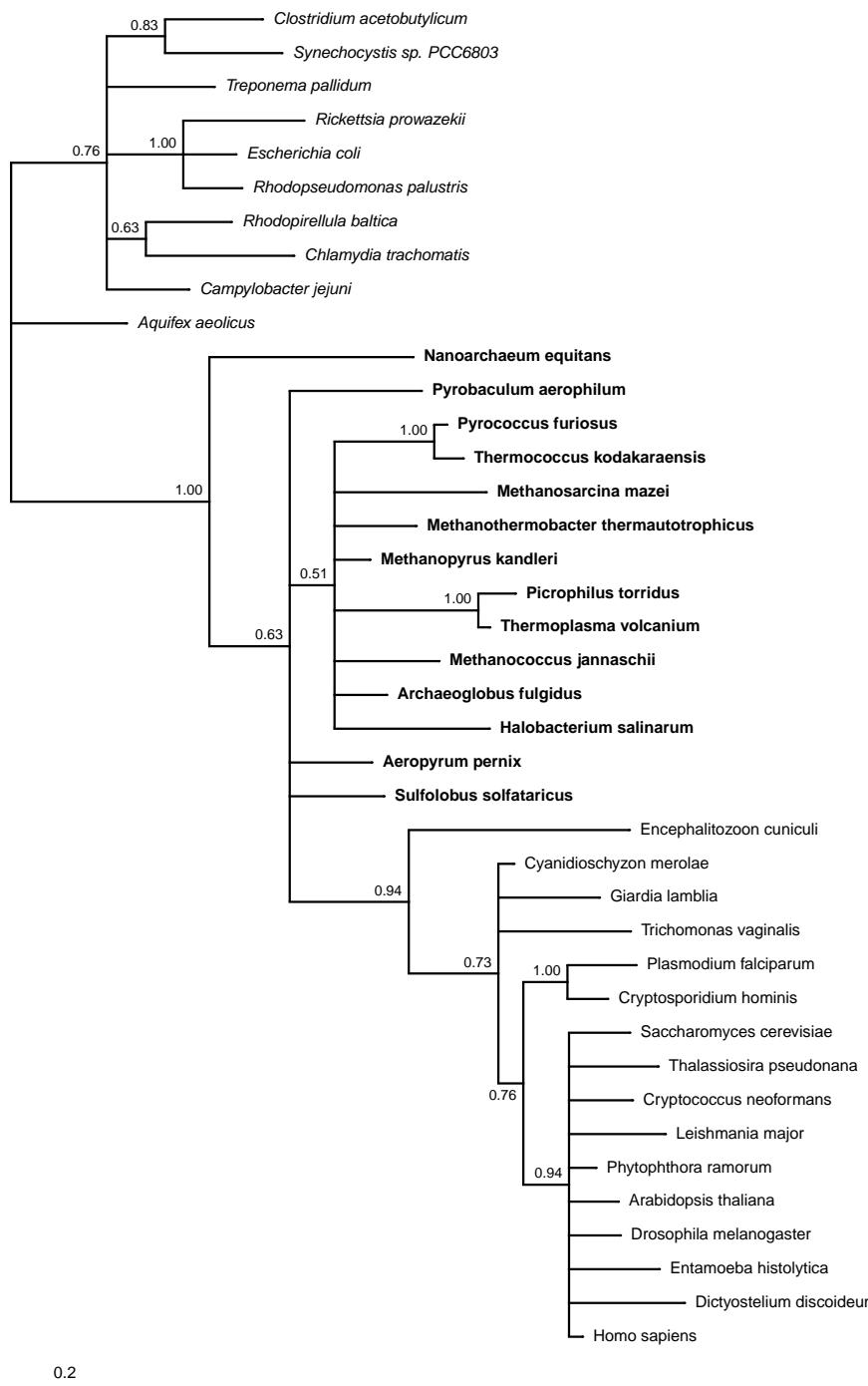
**Fig. S14:** 60S ribosomal protein L11 (L5) – nTax = 40, nChar = 109 Substitution model: WAG+Γ+1CV  
Composition homogeneity test P value = 0.4960



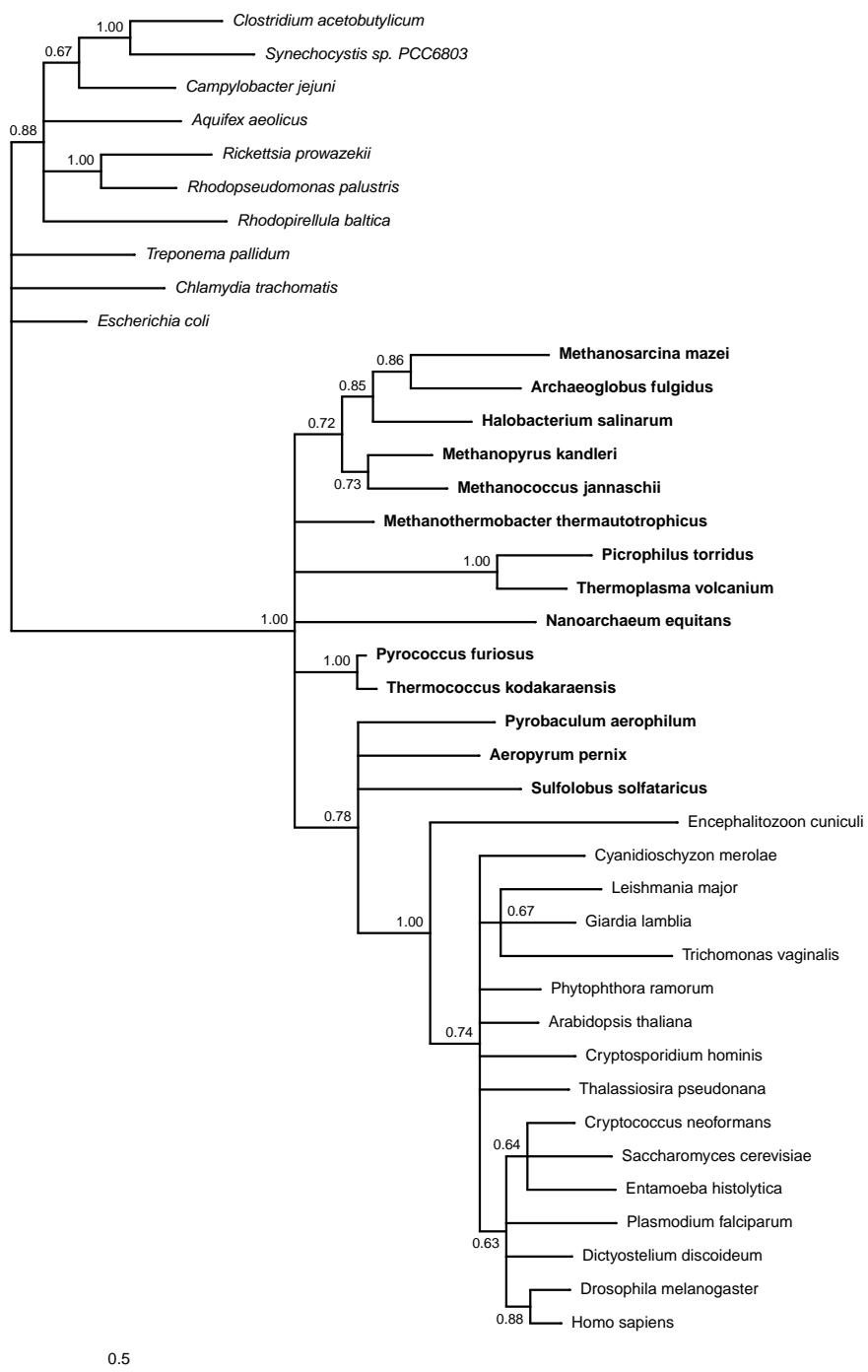
**Fig. S15:** 60S ribosomal protein L17 (L22) – nTax = 40, nChar = 63 Substitution model: WAG+I+Γ+4CV  
Composition homogeneity test P value = 0.0864



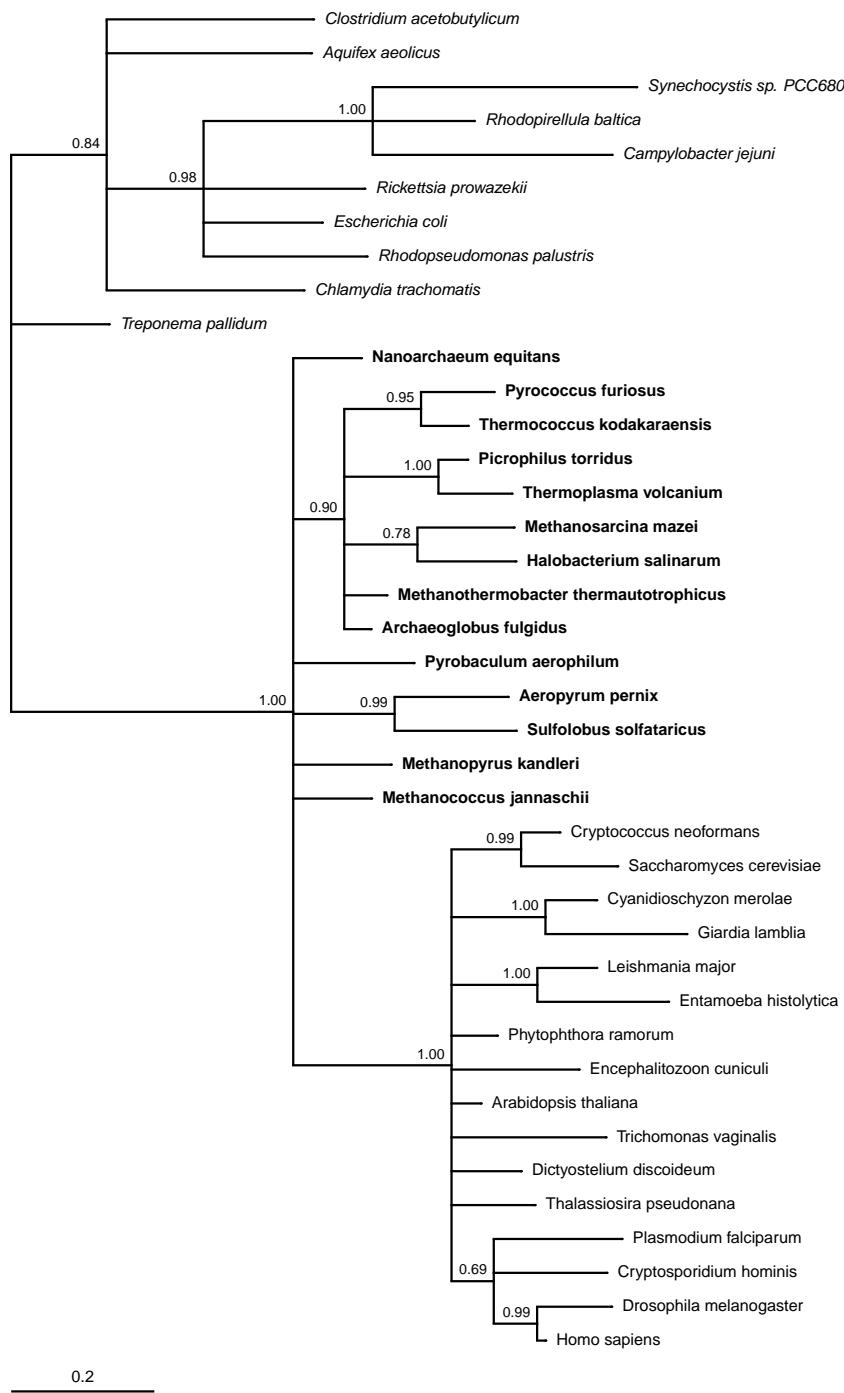
**Fig. S16:** 60S ribosomal protein L12 (L11) – nTax = 39, nChar = 99 Substitution model: WAG+Γ+3CV  
Composition homogeneity test P value = 0.1364



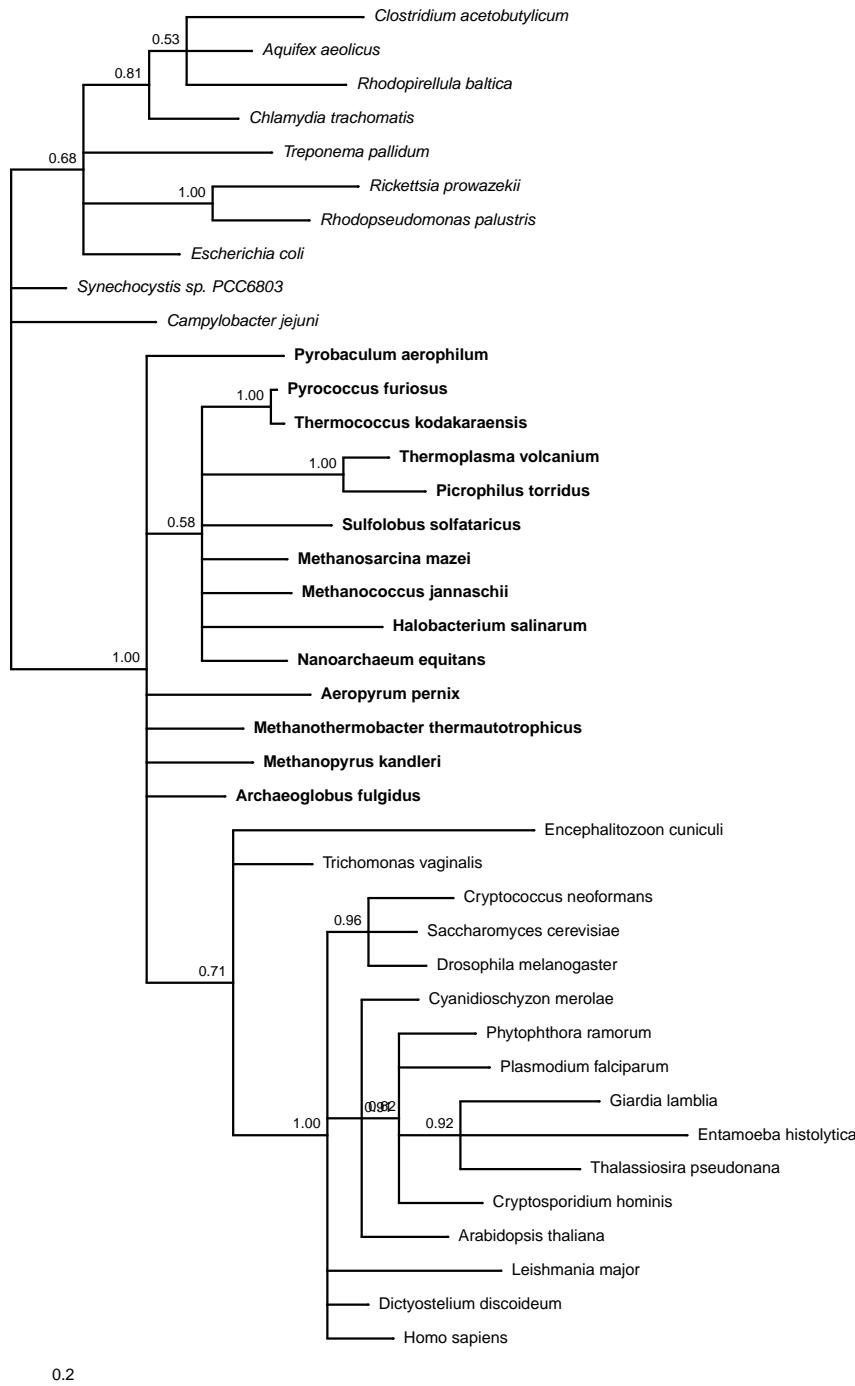
**Fig. S17:** 60S ribosomal protein L23 (L14) – nTax = 40, nChar = 90 Substitution model: WAG+I+Γ+2CV  
Composition homogeneity test P value = 0.1909



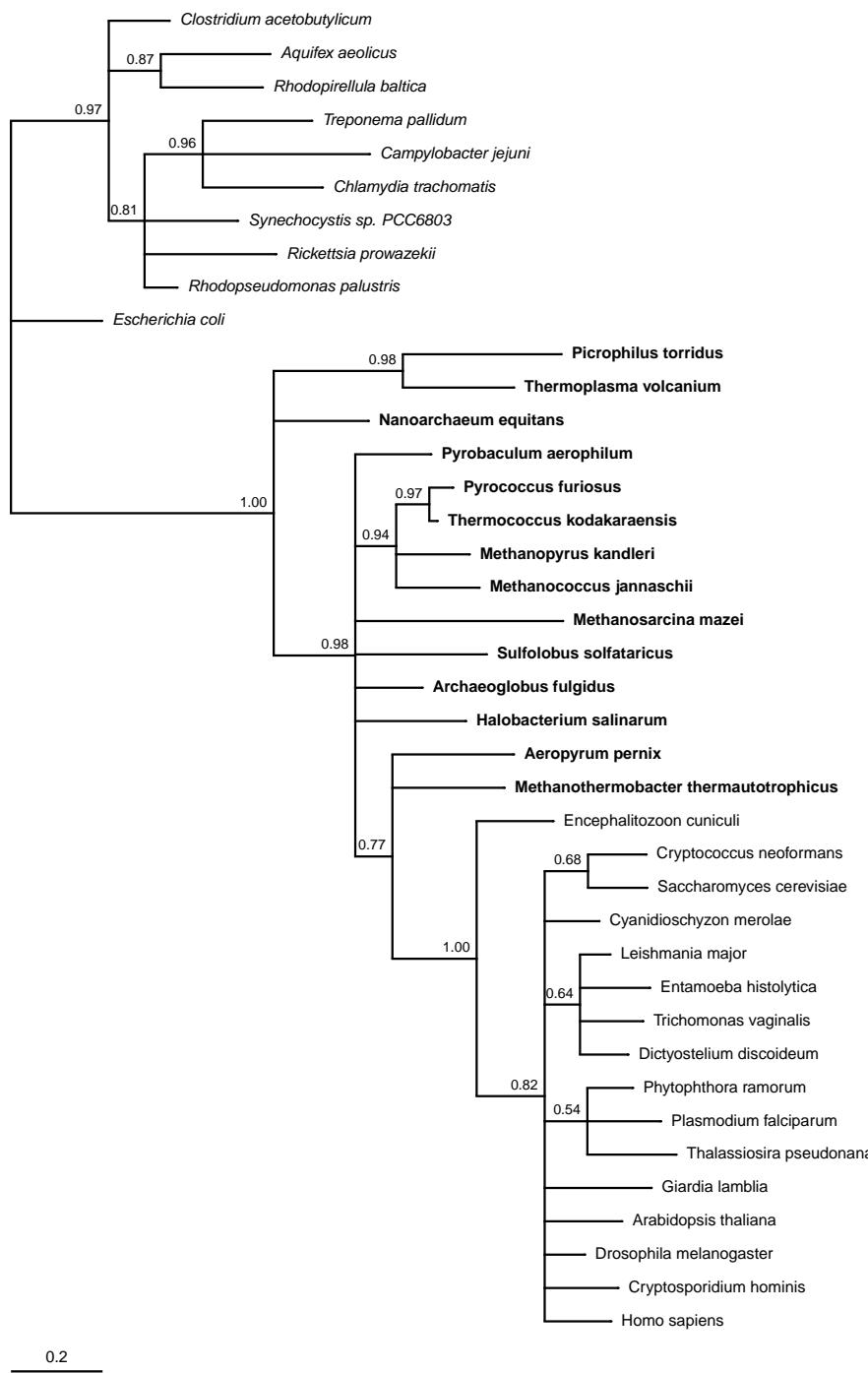
**Fig. S18:** 60S ribosomal protein L8 (L2) – nTax = 40, nChar = 159 Substitution model: WAG+I+Γ+2CV  
Composition homogeneity test P value = 0.1384



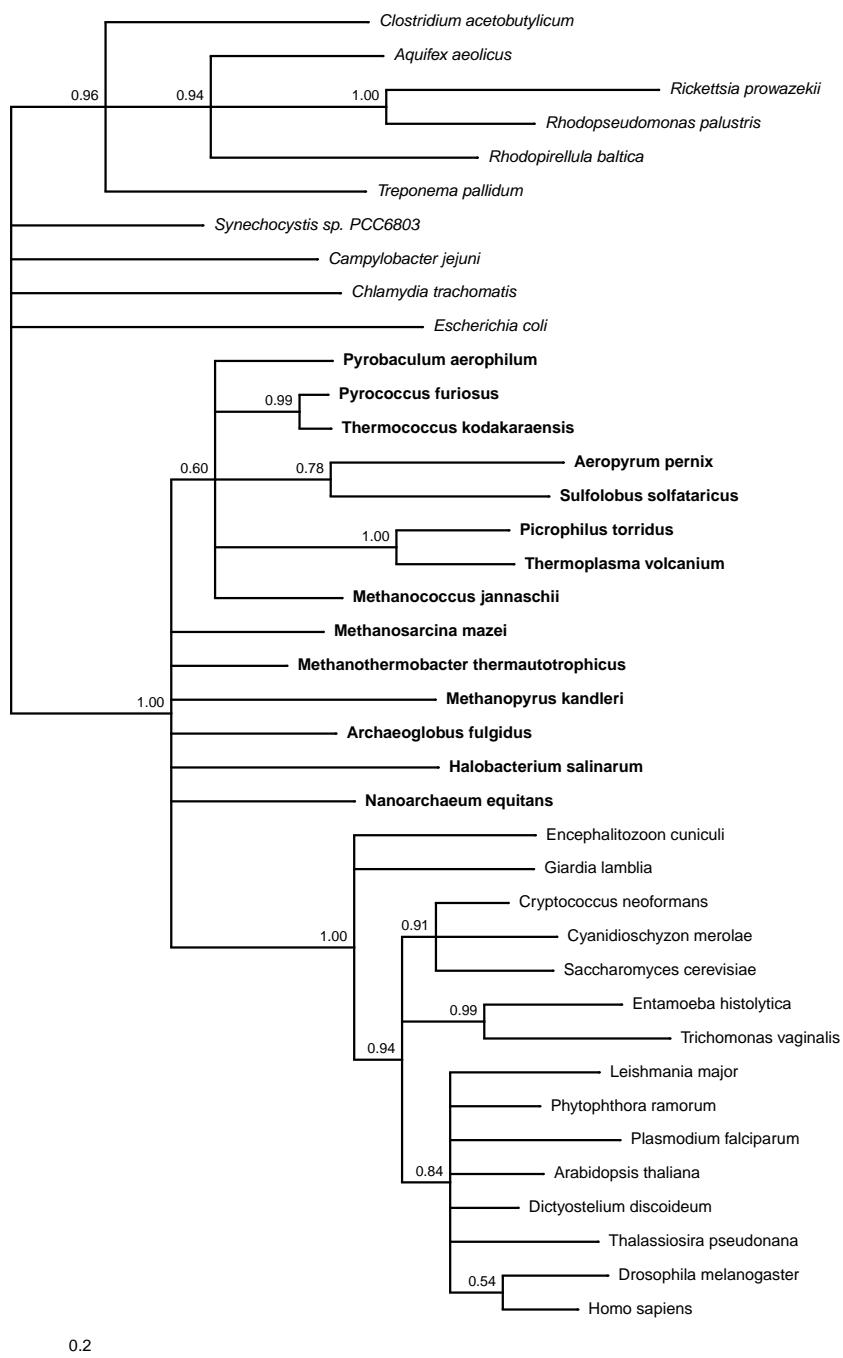
**Fig. S19:** 60S ribosomal protein L3 – nTax = 40, nChar = 63 Substitution model: WAG+I+Γ+1CV Composition homogeneity test P value = 0.0914



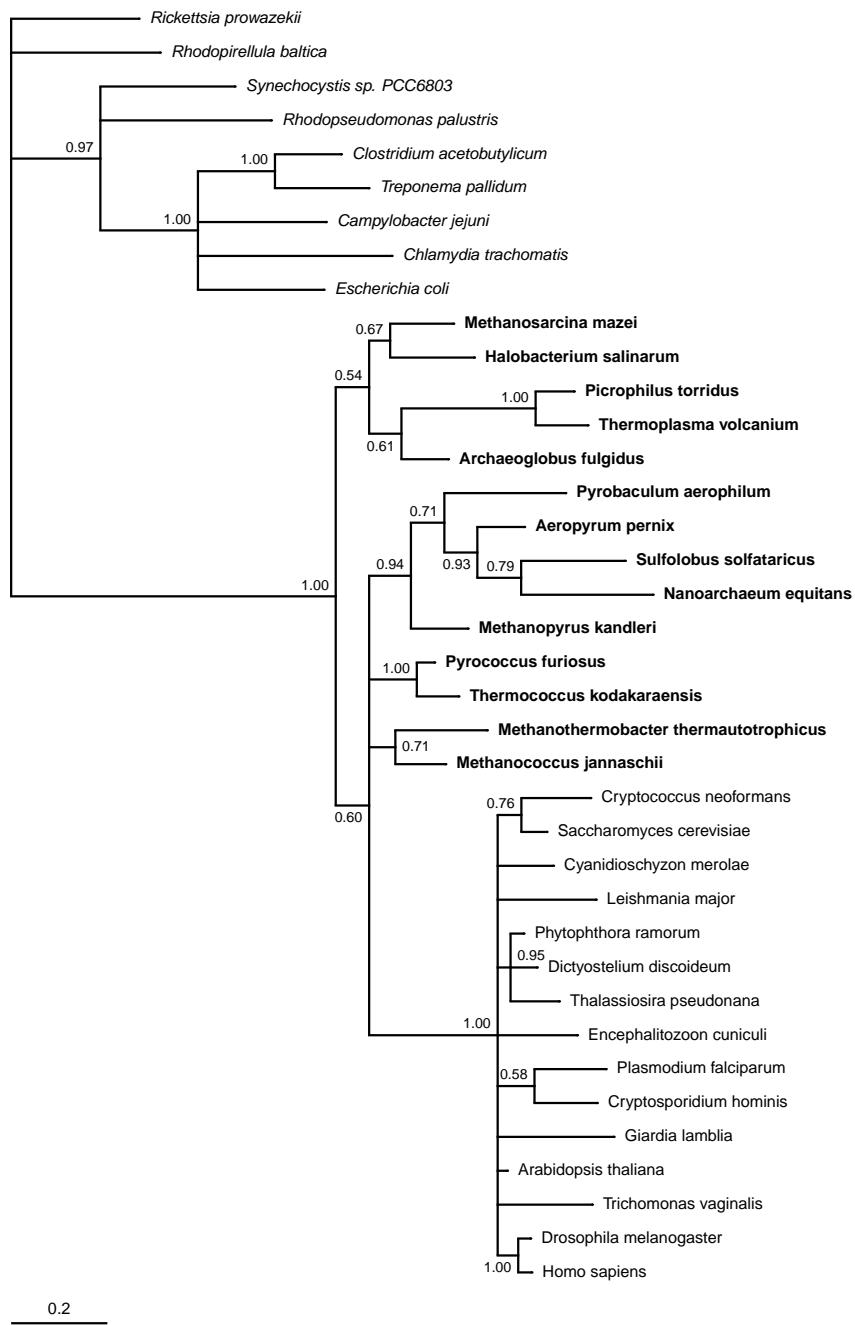
**Fig. S20:** 60S ribosomal protein L13A – nTax = 40, nChar = 65 Substitution model: WAG+I+Γ+1CV Composition homogeneity test P value = 0.1177



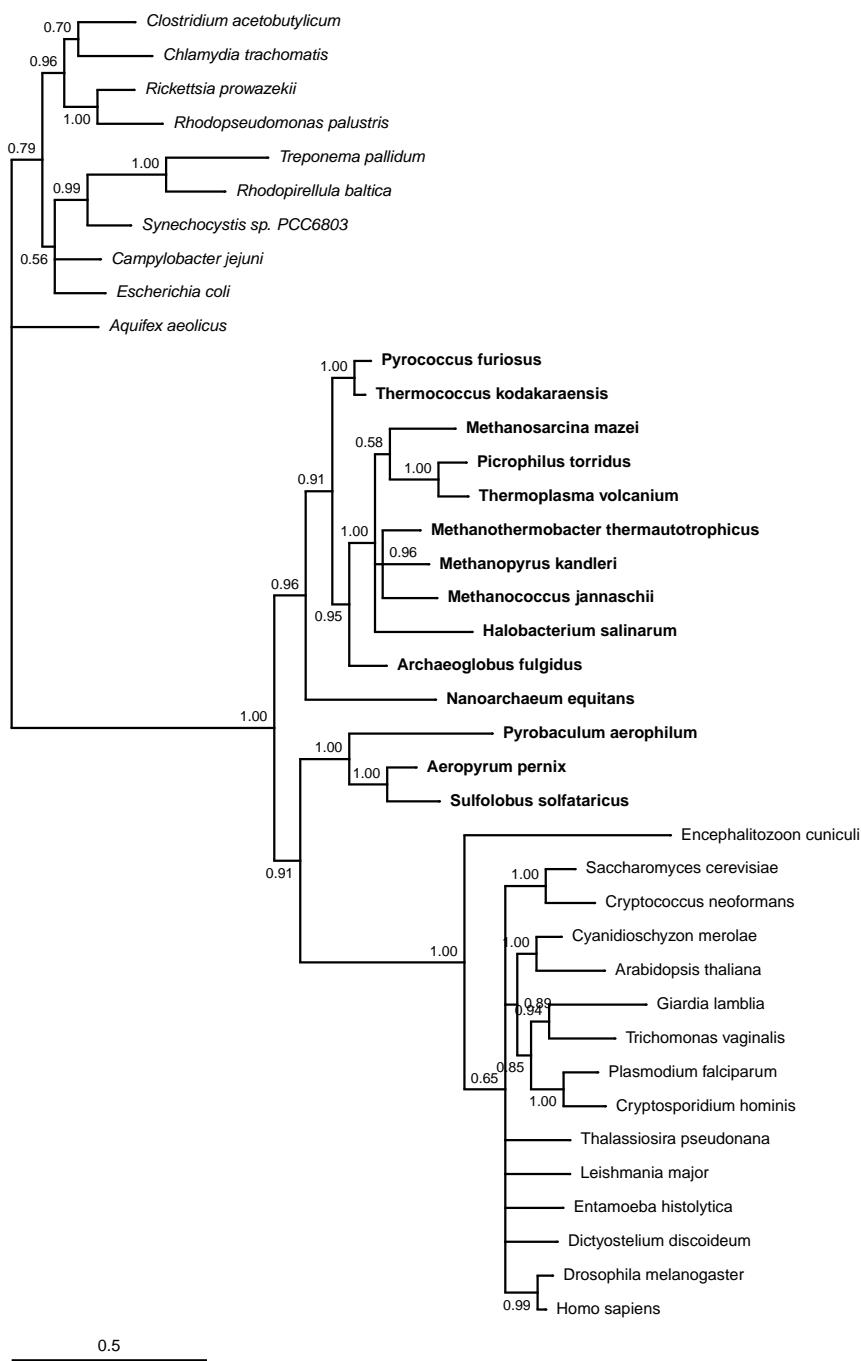
**Fig. S21:** 60S ribosomal protein L5 (L18) – nTax = 40, nChar = 64 Substitution model: WAG+I+Γ+3CV  
Composition homogeneity test P value = 0.0627



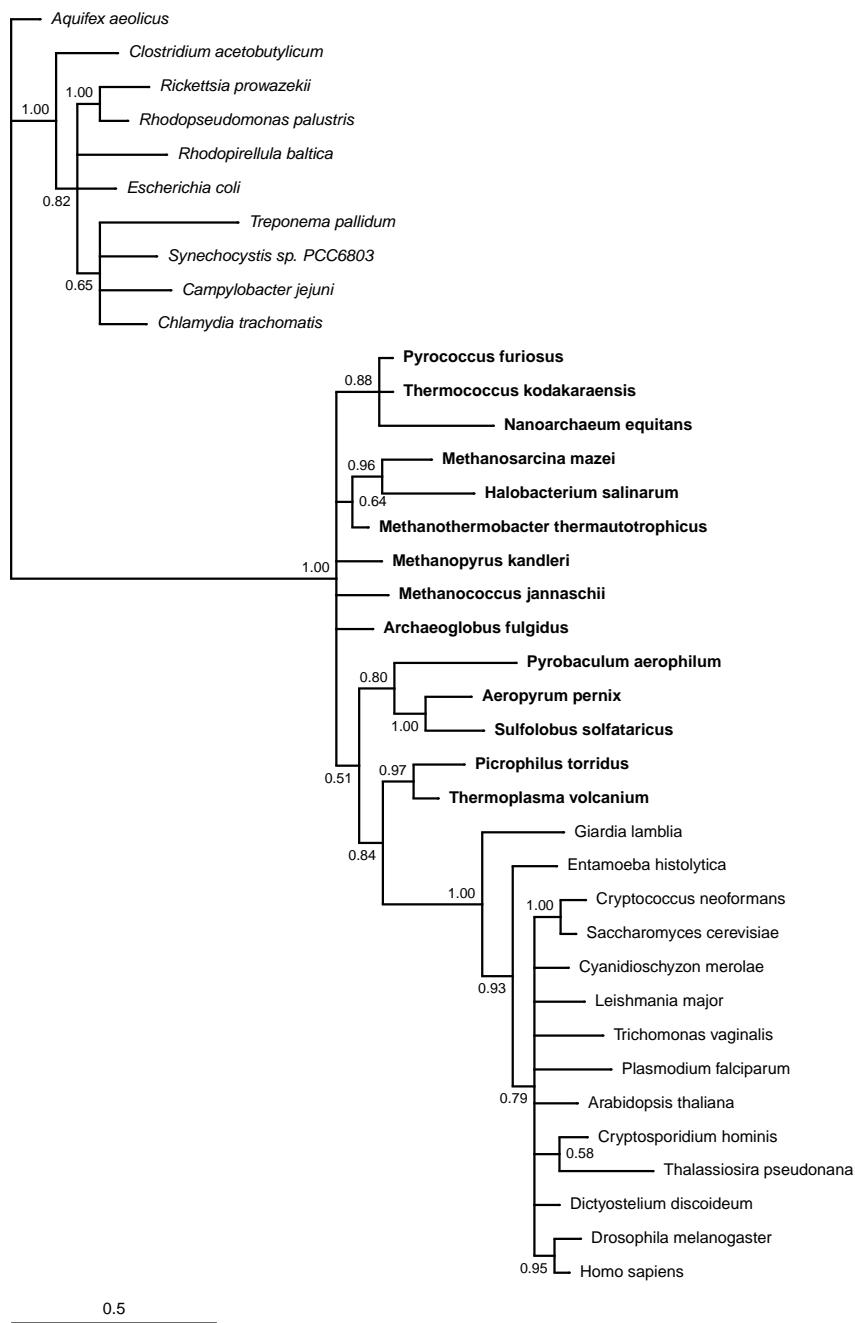
**Fig. S22:** 60S ribosomal protein L4 – nTax = 39, nChar = 68 Substitution model: WAG+I+Γ+3CV Composition homogeneity test P value = 0.1123



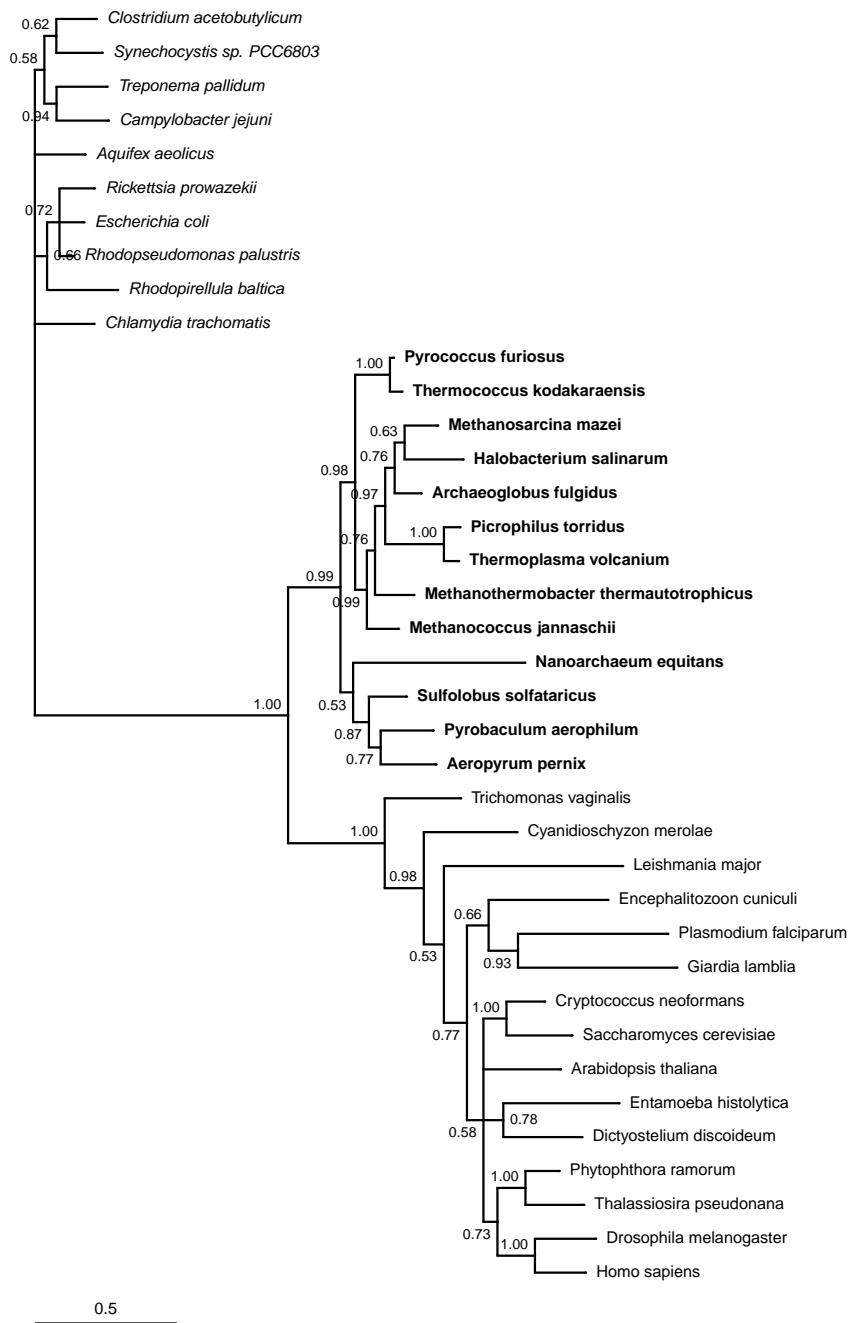
**Fig. S23:** ATP-binding cassette sub-family E (RLI) member 1 – nTax = 38, nChar = 134 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1272



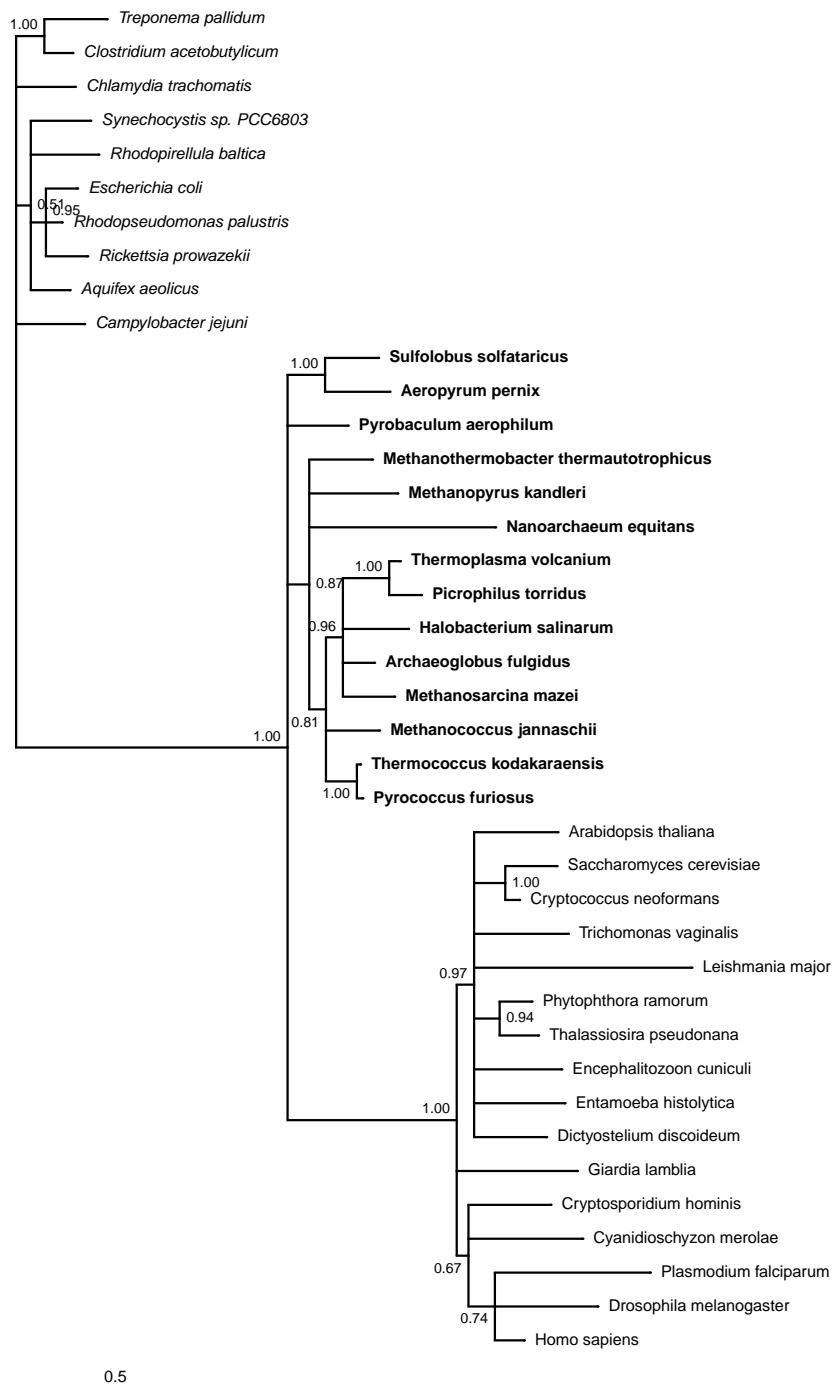
**Fig. S24:** Elongation factor 2 (EF-G) – nTax = 39, nChar = 271 Substitution model: WAG+I+Γ+3CV Composition homogeneity test P value = 0.0818



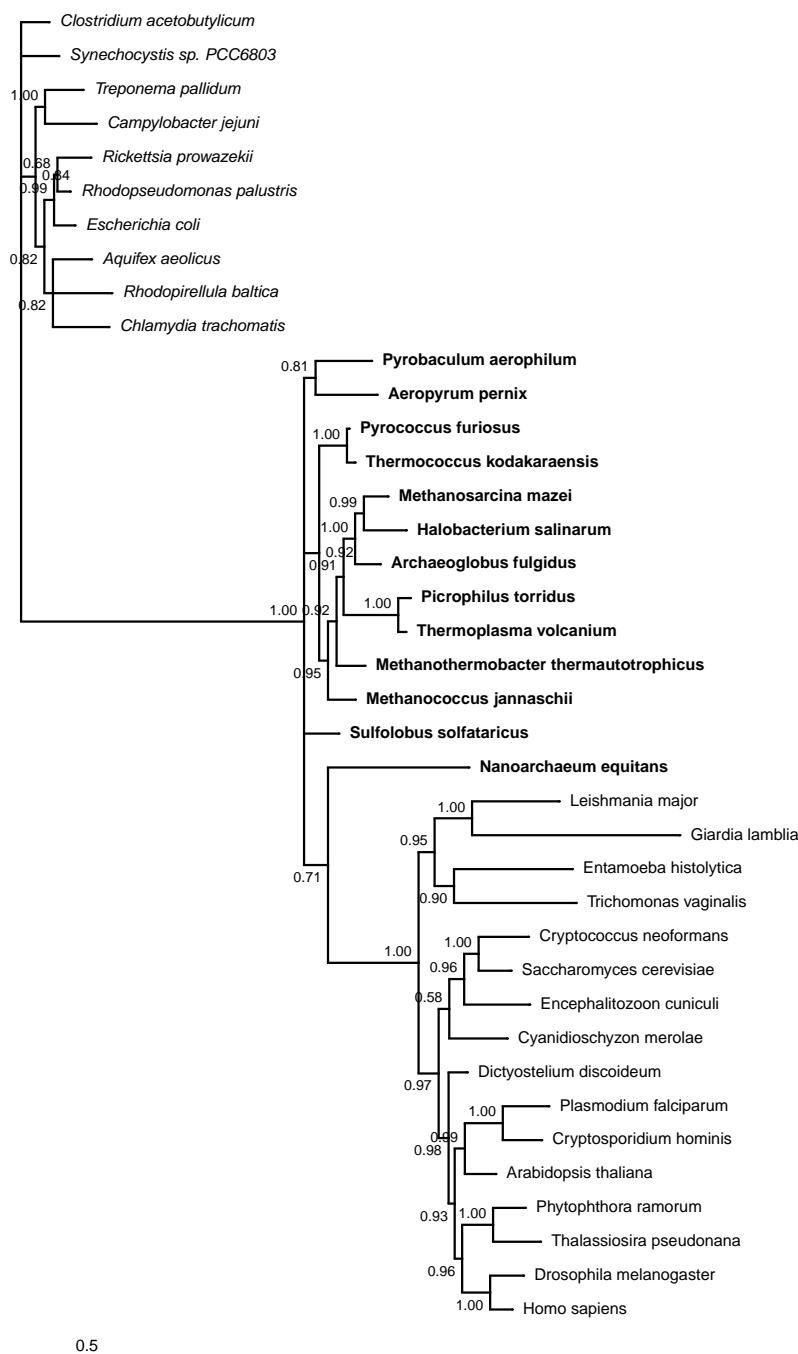
**Fig. S25:** Eukaryotic translation elongation factor 1 $\alpha$  (EF Tu) – nTax = 38, nChar = 202 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.8595



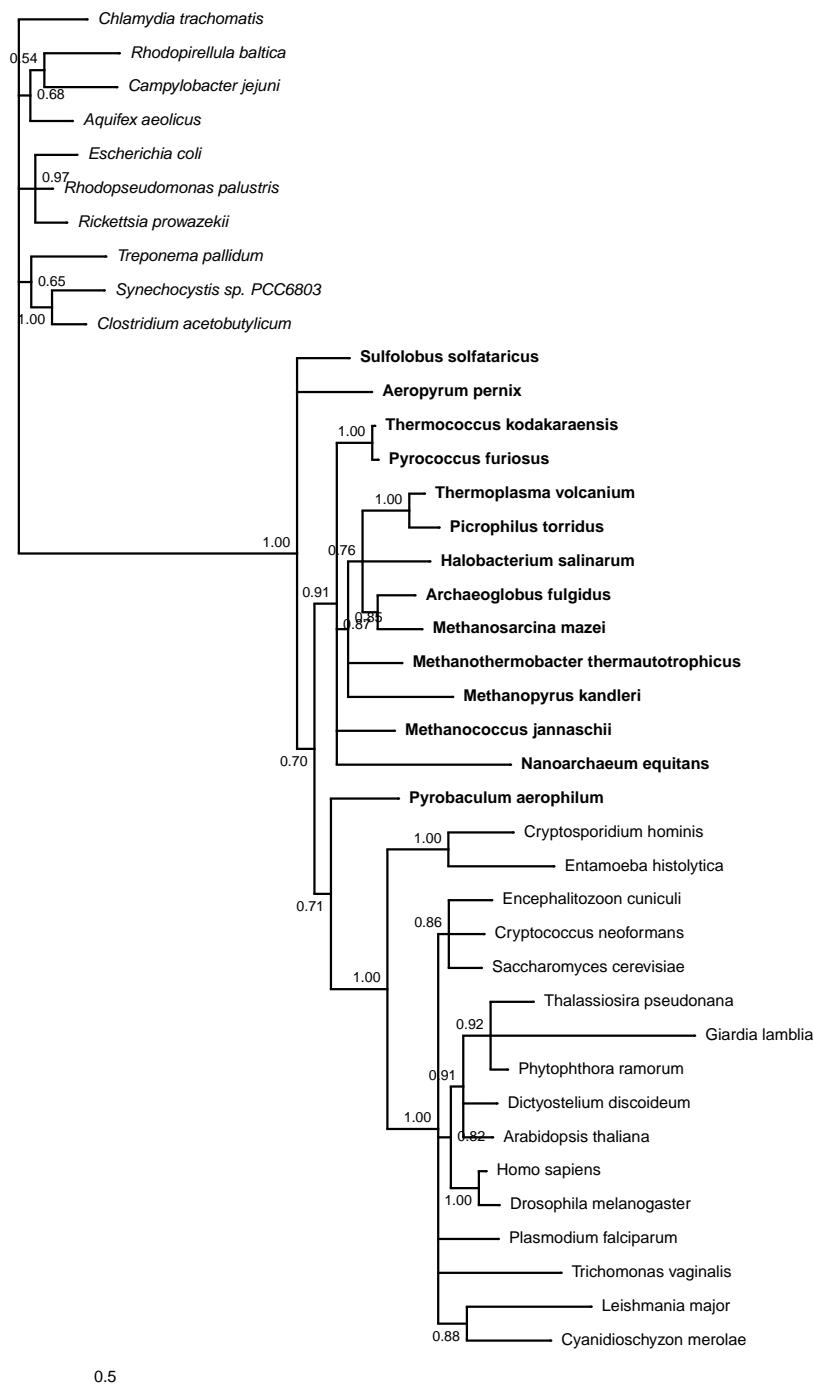
**Fig. S26:** RNA polymerase I RPA1 – nTax = 38, nChar = 332 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1482



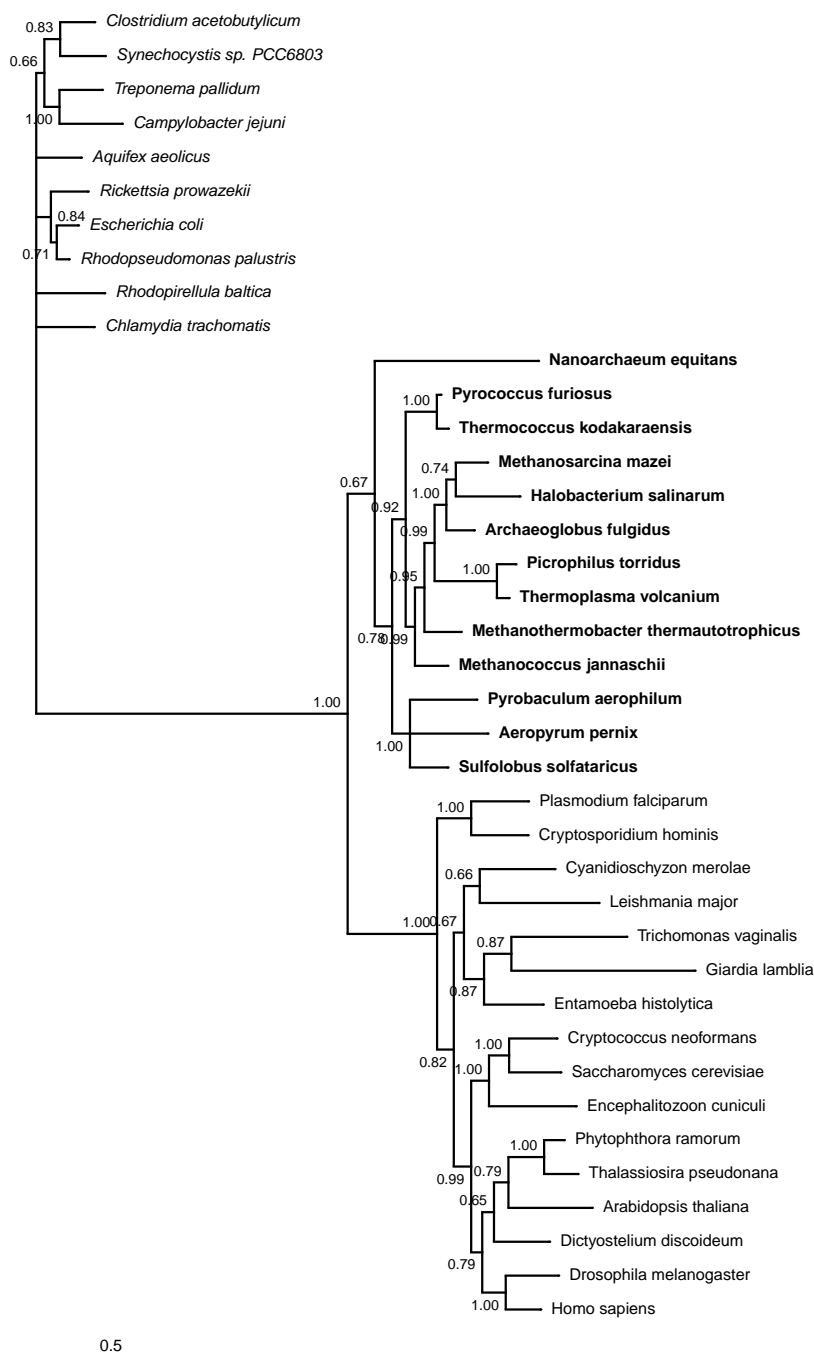
**Fig. S27:** RNA polymerase I RPA2 – nTax = 40, nChar = 222 Substitution model: WAG+I+Γ+1CV Composition homogeneity test P value = 0.3338



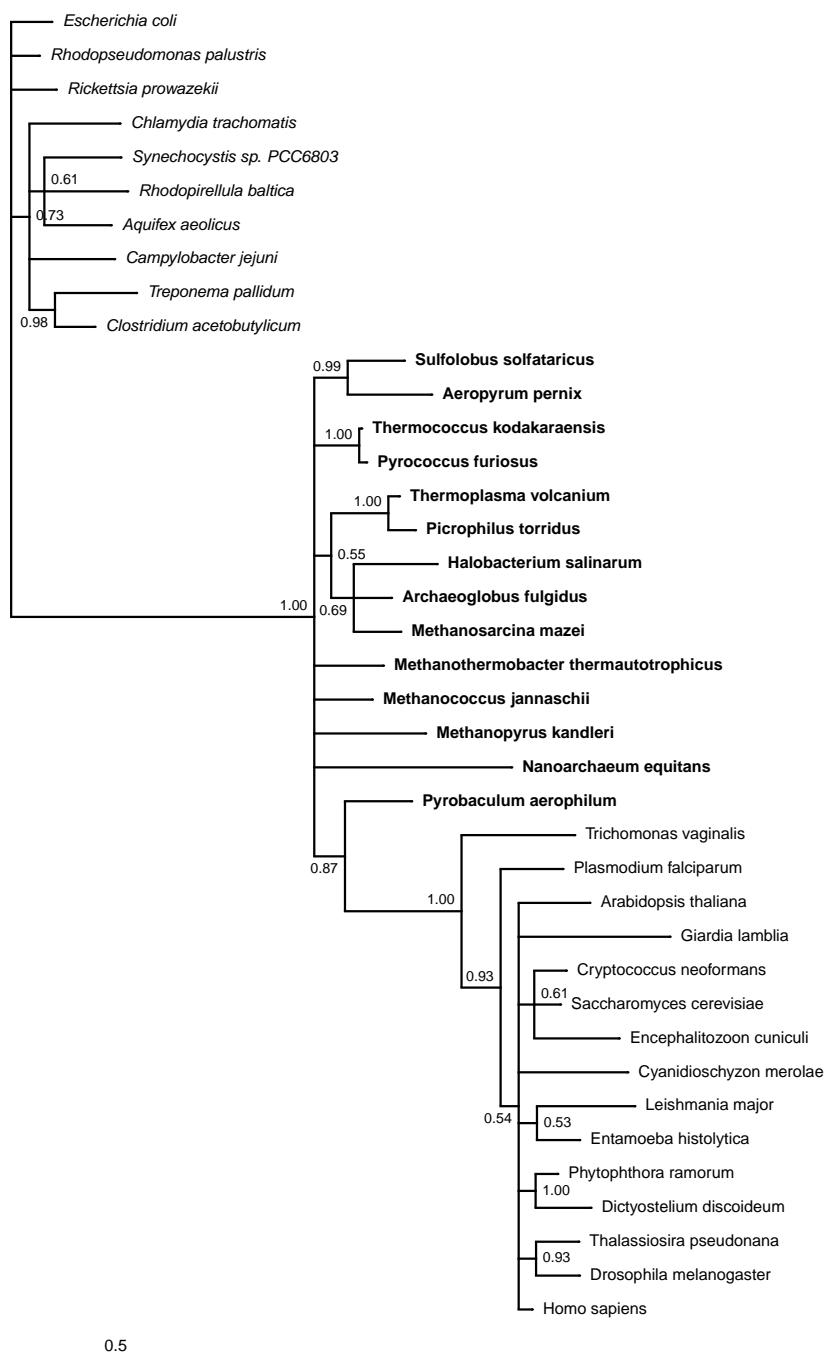
**Fig. S28:** RNA polymerase II RPB1 – nTax = 39, nChar = 432 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.2233



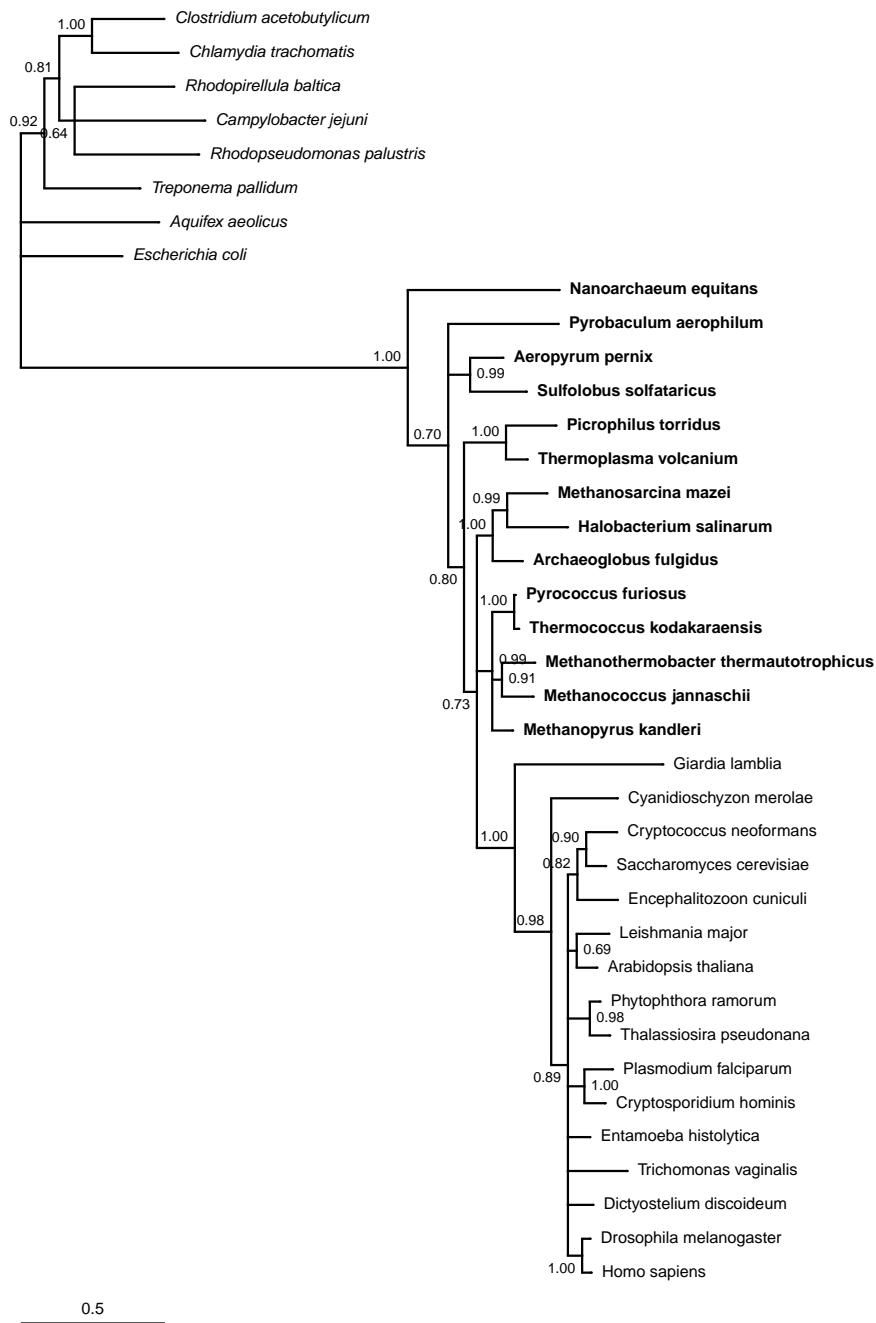
**Fig. S29:** RNA polymerase II RPB2 – nTax = 40, nChar = 313 Substitution model: WAG+I+Γ+1CV Composition homogeneity test P value = 0.1280



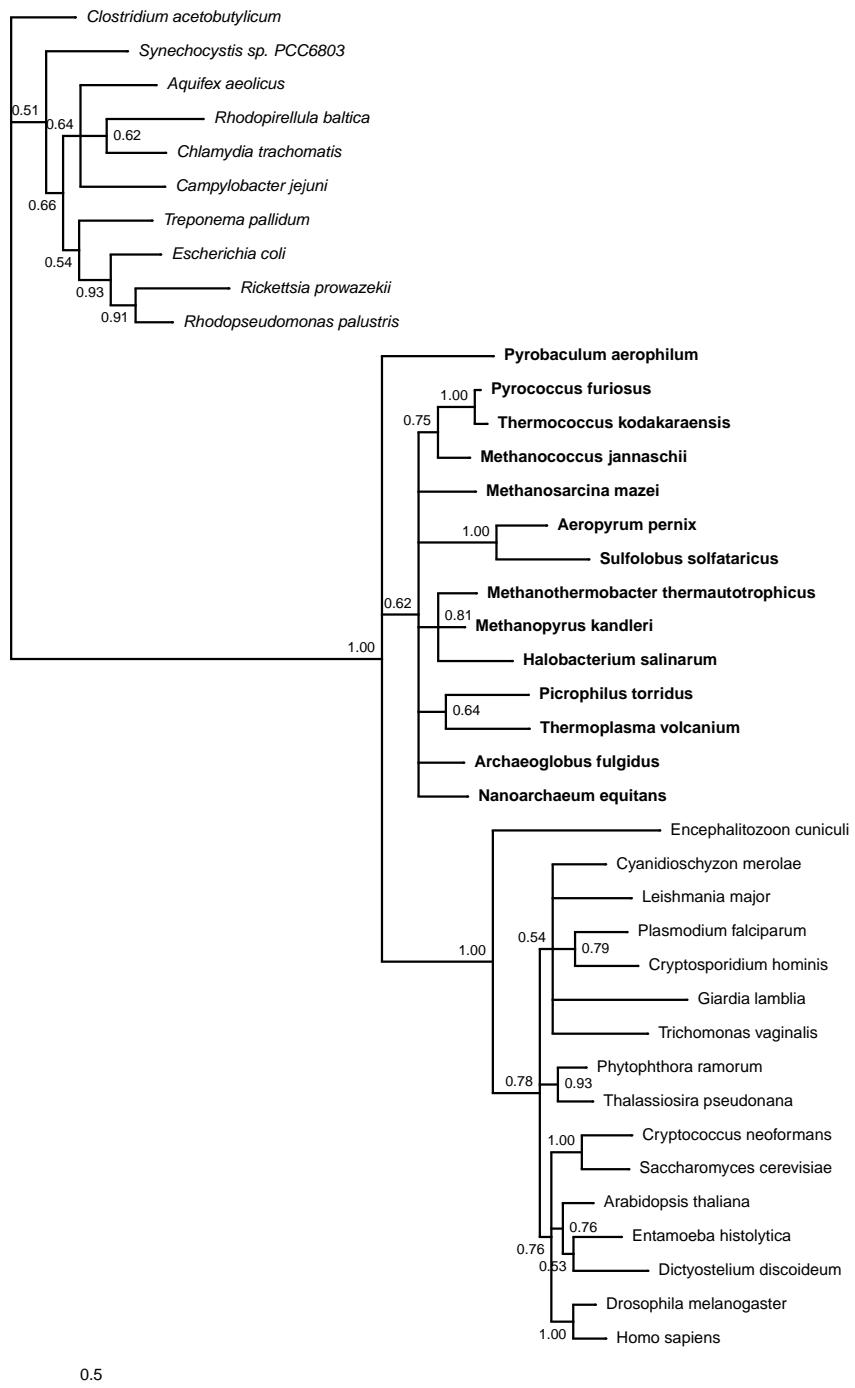
**Fig. S30:** RNA polymerase III RPC1 – nTax = 39, nChar = 377 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1485



**Fig. S31:** RNA polymerase III RPC2 – nTax = 39, nChar = 267 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1165



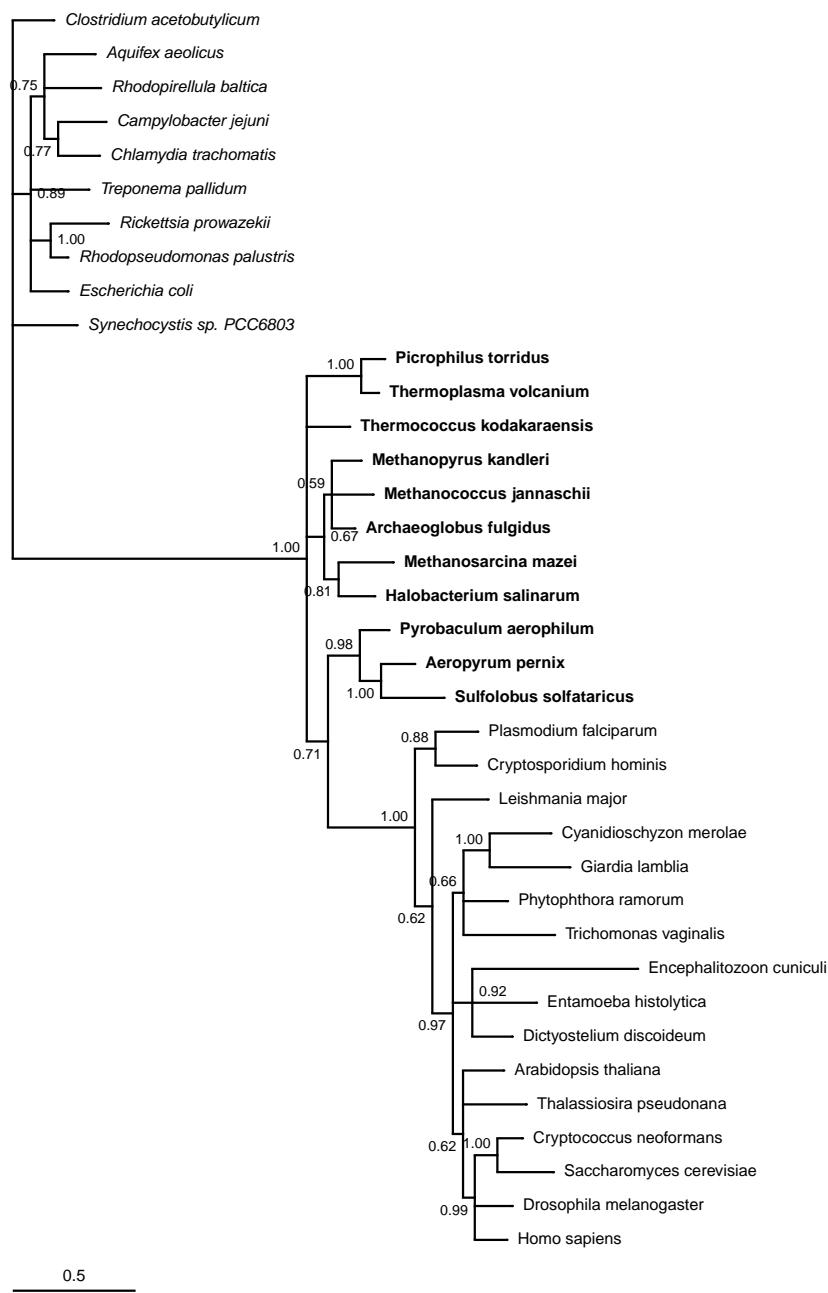
**Fig. S32:** V-type ATPase V1 subunit B – nTax = 38, nChar = 285 Substitution model: WAG+I+Γ+4CV Composition homogeneity test P value = 0.1150



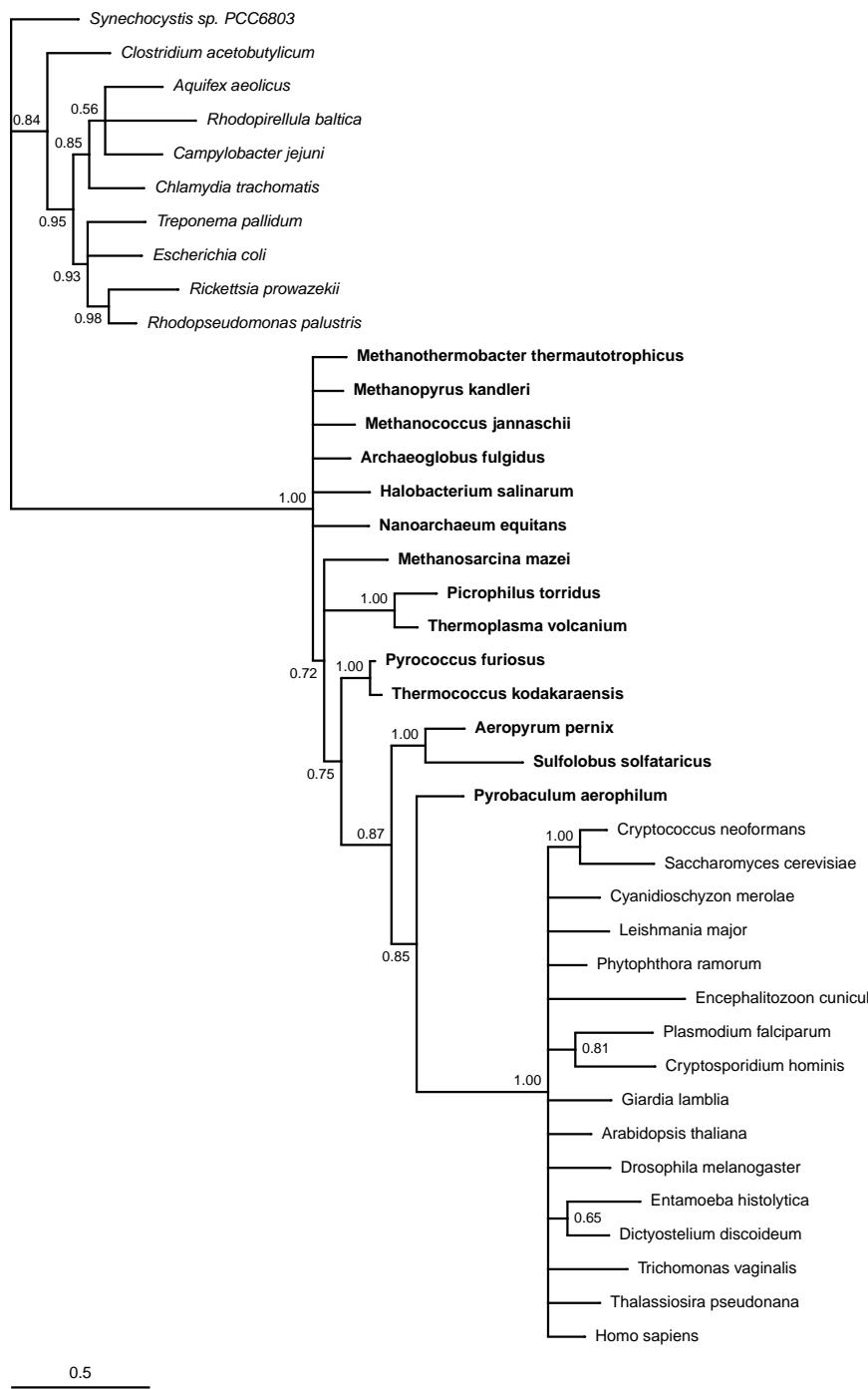
**Fig. S33:** Chaperonin containing TCP1 subunit 1 ( $\alpha$ ) – nTax = 40, nChar = 202 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1227



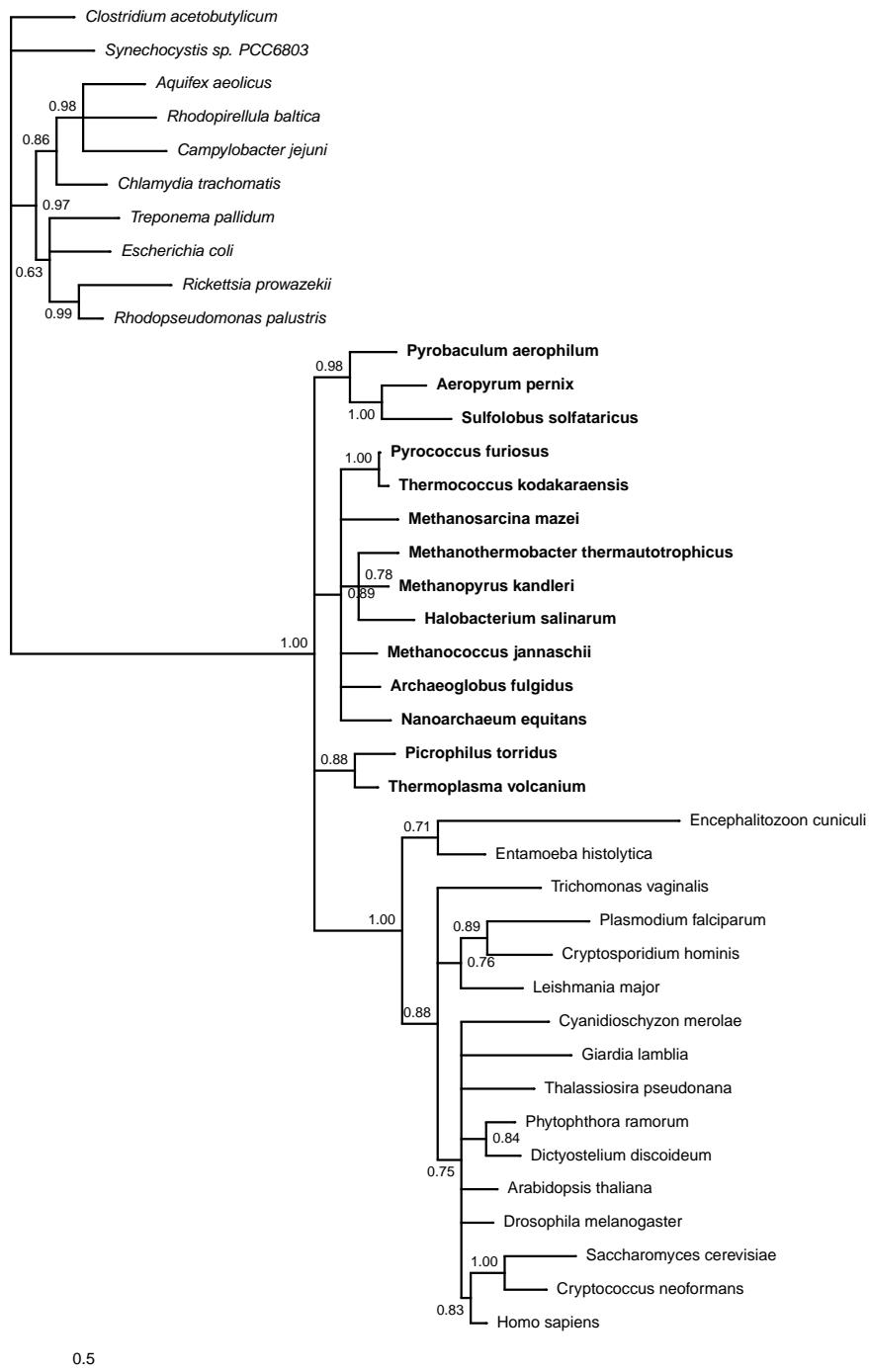
**Fig. S34:** Chaperonin containing TCP1 subunit 3 ( $\gamma$ ) – nTax = 40, nChar = 175 Substitution model: WAG+I+ $\Gamma$ +2CV Composition homogeneity test P value = 0.3257



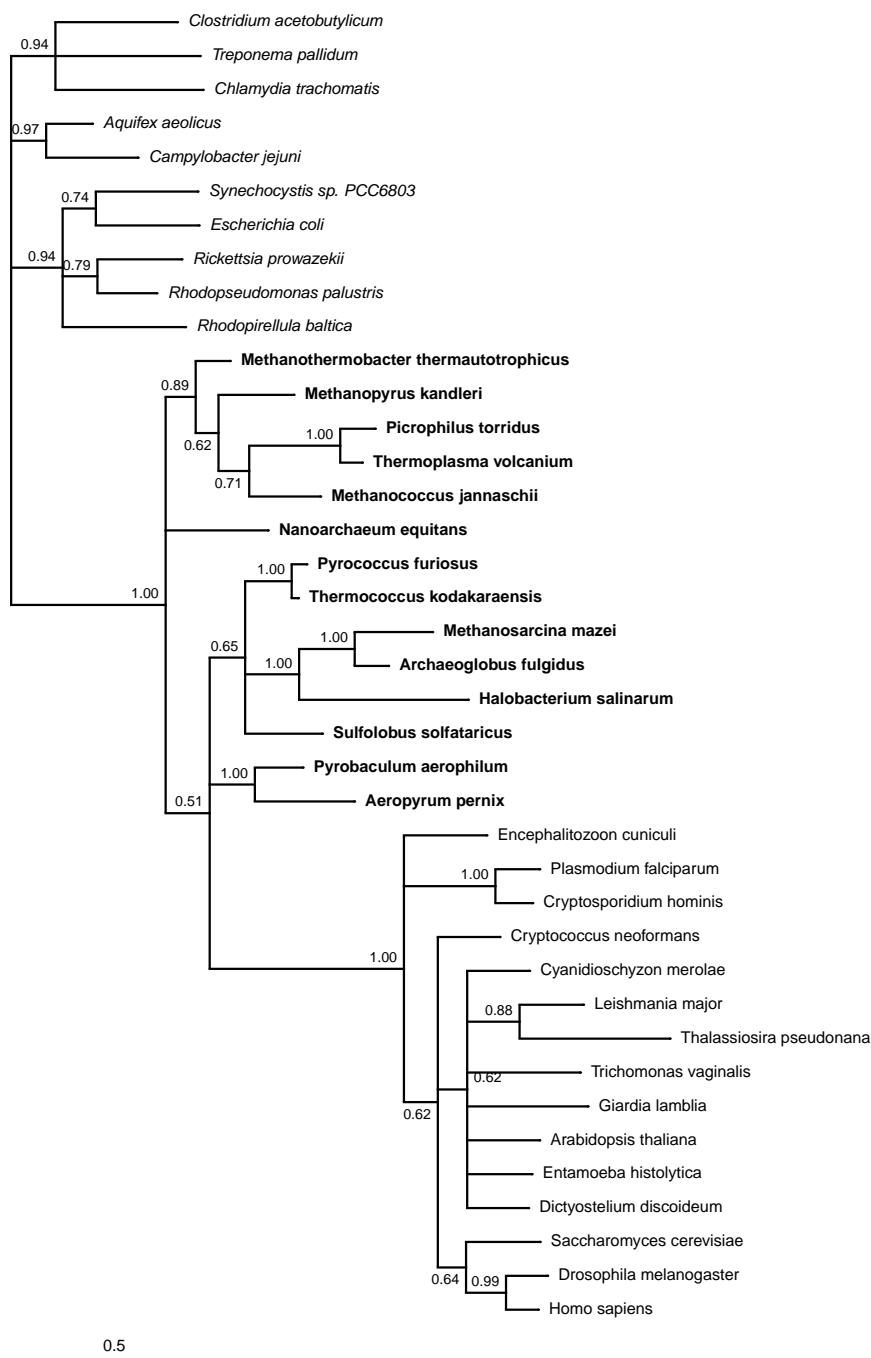
**Fig. S35:** Chaperonin containing TCP1 subunit 4 ( $\delta$ ) – nTax = 37, nChar = 218 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.0520



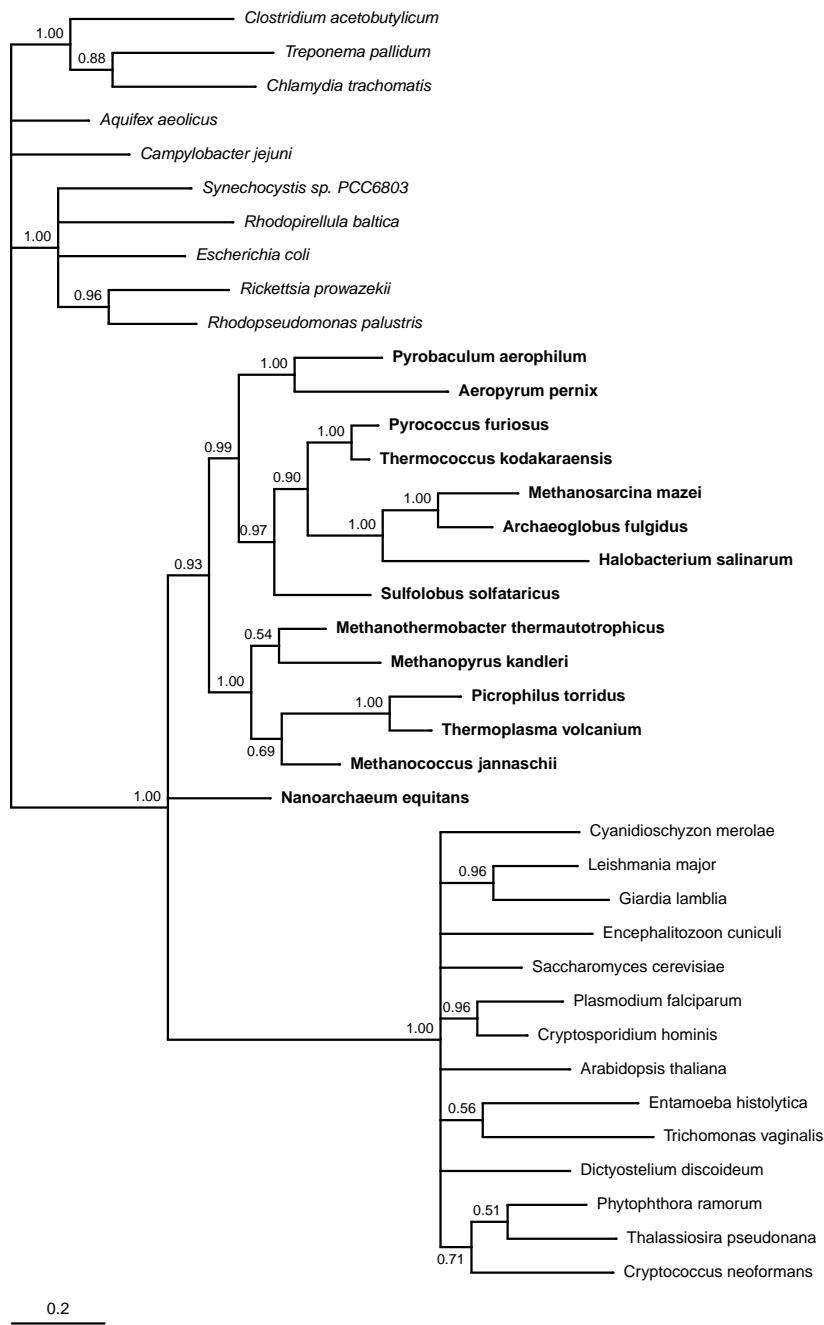
**Fig. S36:** Chaperonin containing TCP1 subunit 5 ( $\epsilon$ ) – nTax = 40, nChar = 187 Substitution model: WAG+Γ+2CV Composition homogeneity test P value = 0.0807



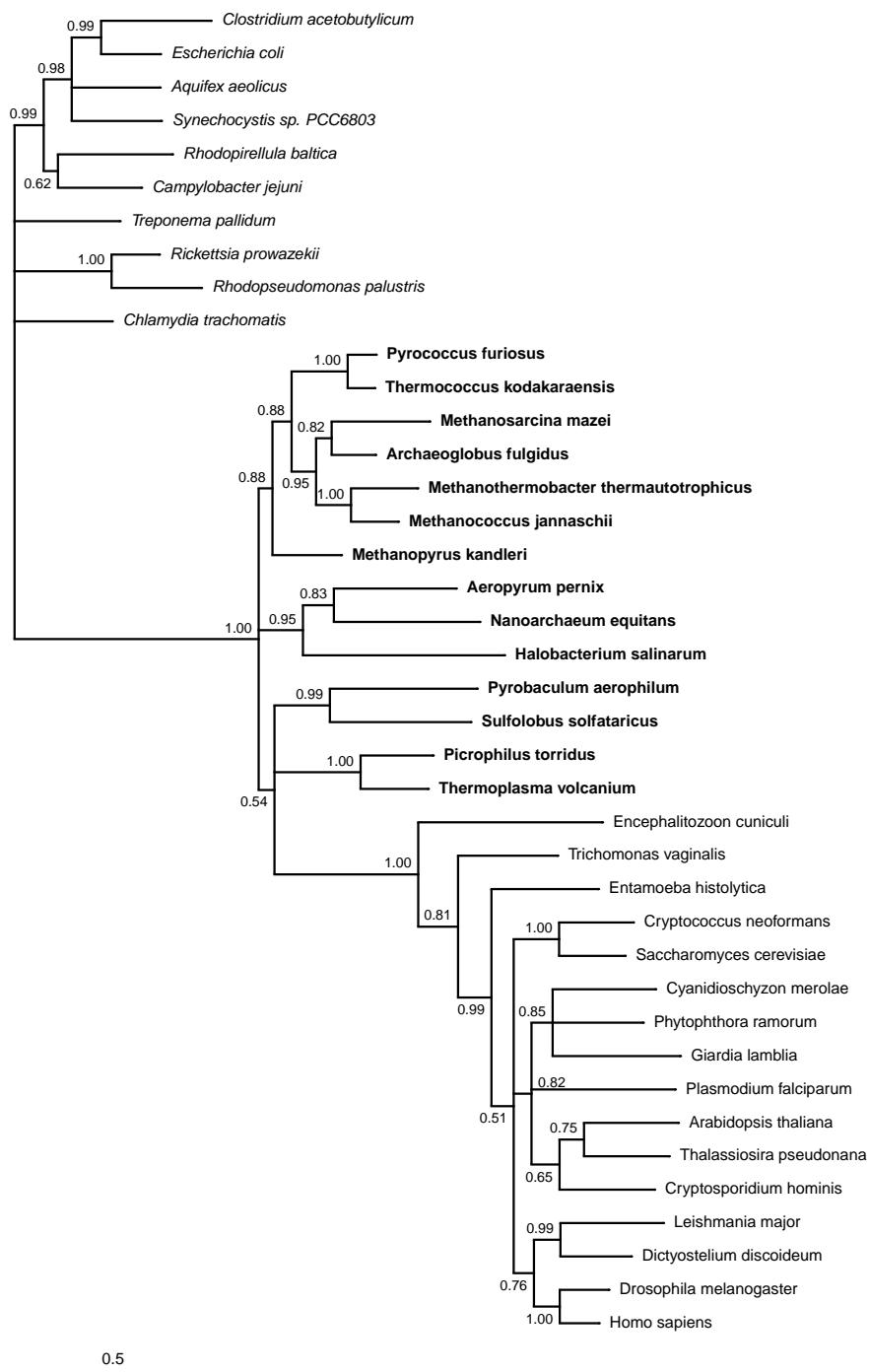
**Fig. S37:** Chaperonin containing TCP1 subunit 7 ( $\eta$ ) – nTax = 40, nChar = 165 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.3351



**Fig. S38:** Glutamine-tRNA ligase – nTax = 39, nChar = 98 Substitution model: WAG+I+Γ+1CV Composition homogeneity test P value = 0.2057



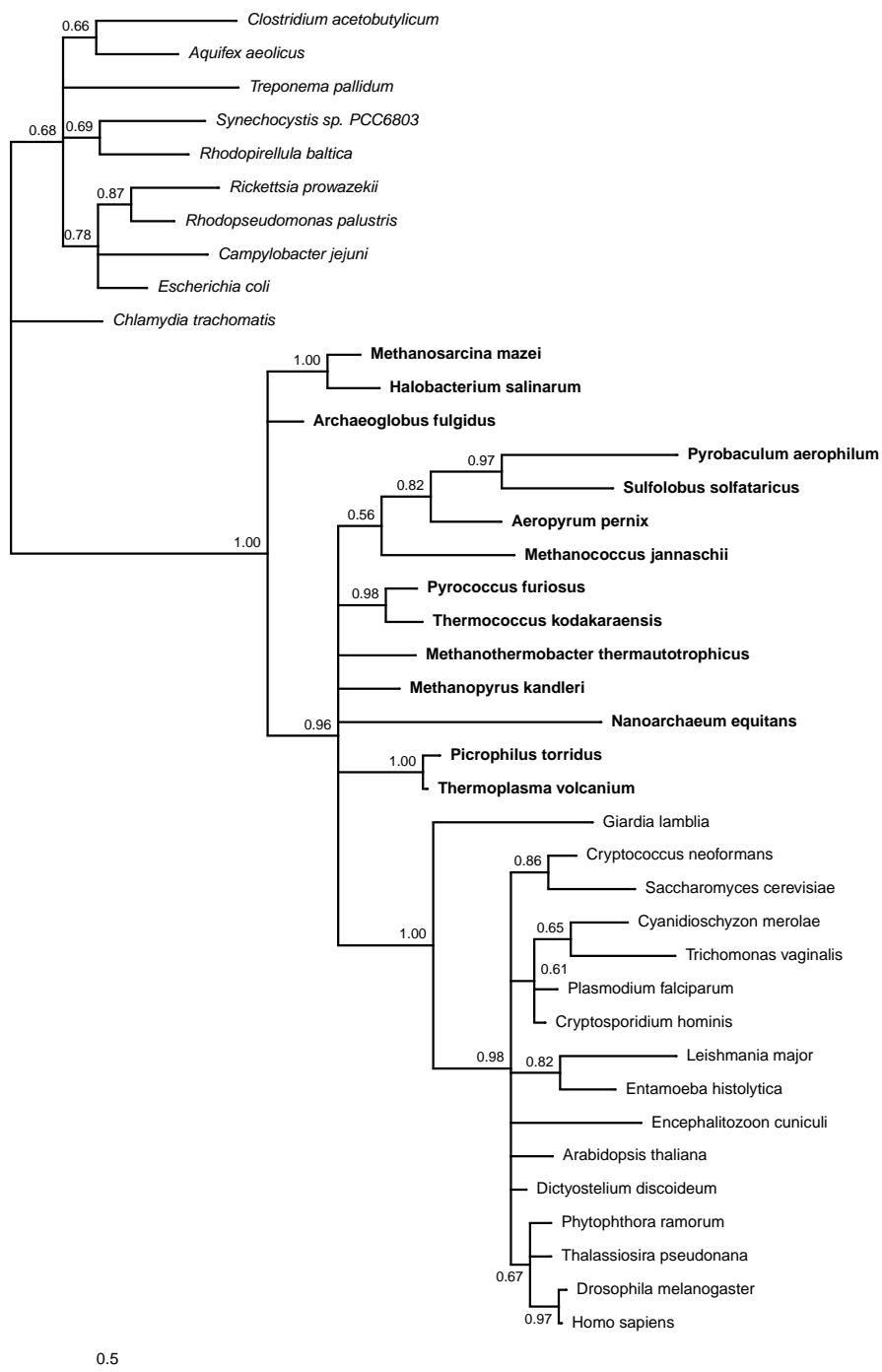
**Fig. S39:** Glutamate-tRNA ligase – nTax = 38, nChar = 137 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.2309



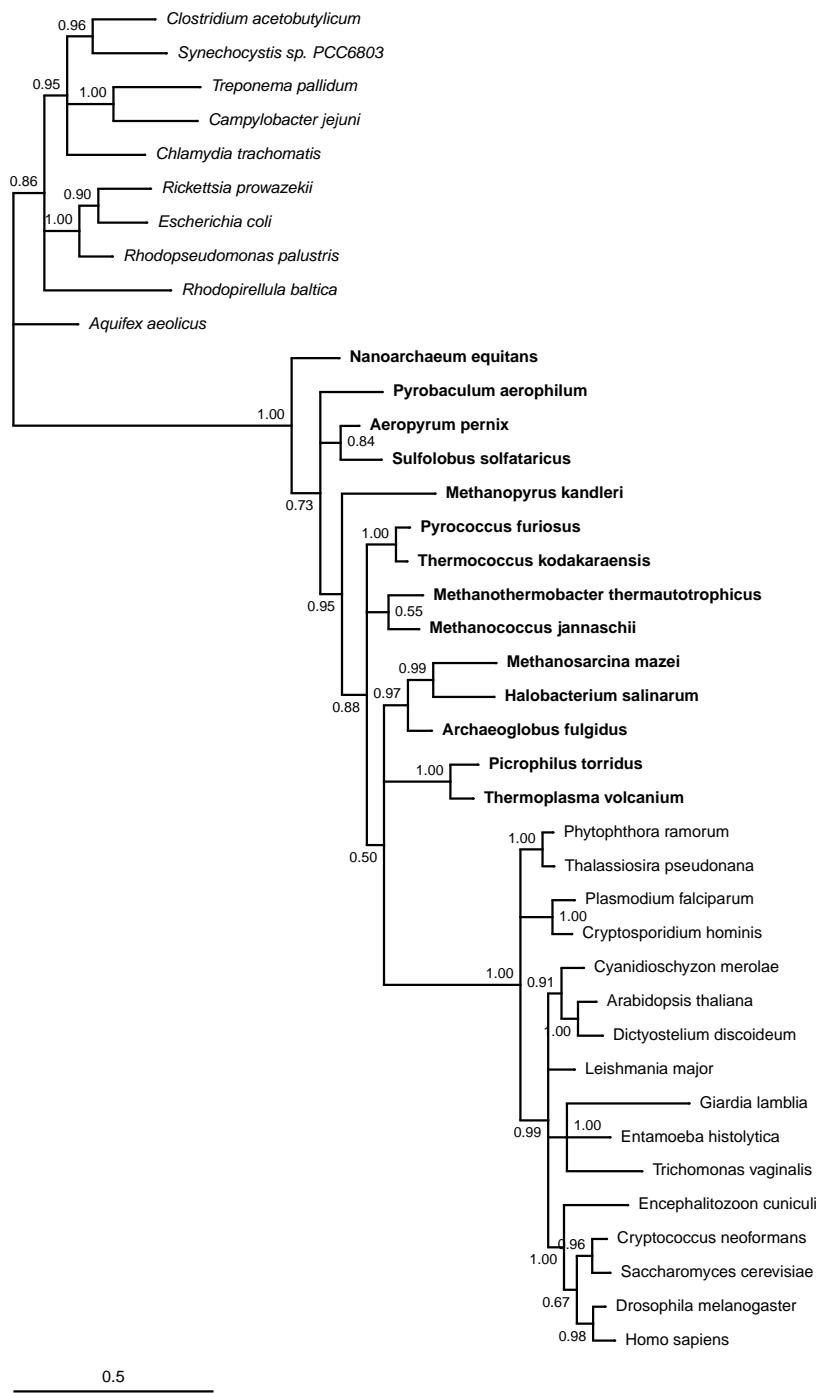
**Fig. S40:** Aspartate-tRNA ligase – nTax = 40, nChar = 223 Substitution model: WAG+I+Γ+4CV Composition homogeneity test P value = 0.0502



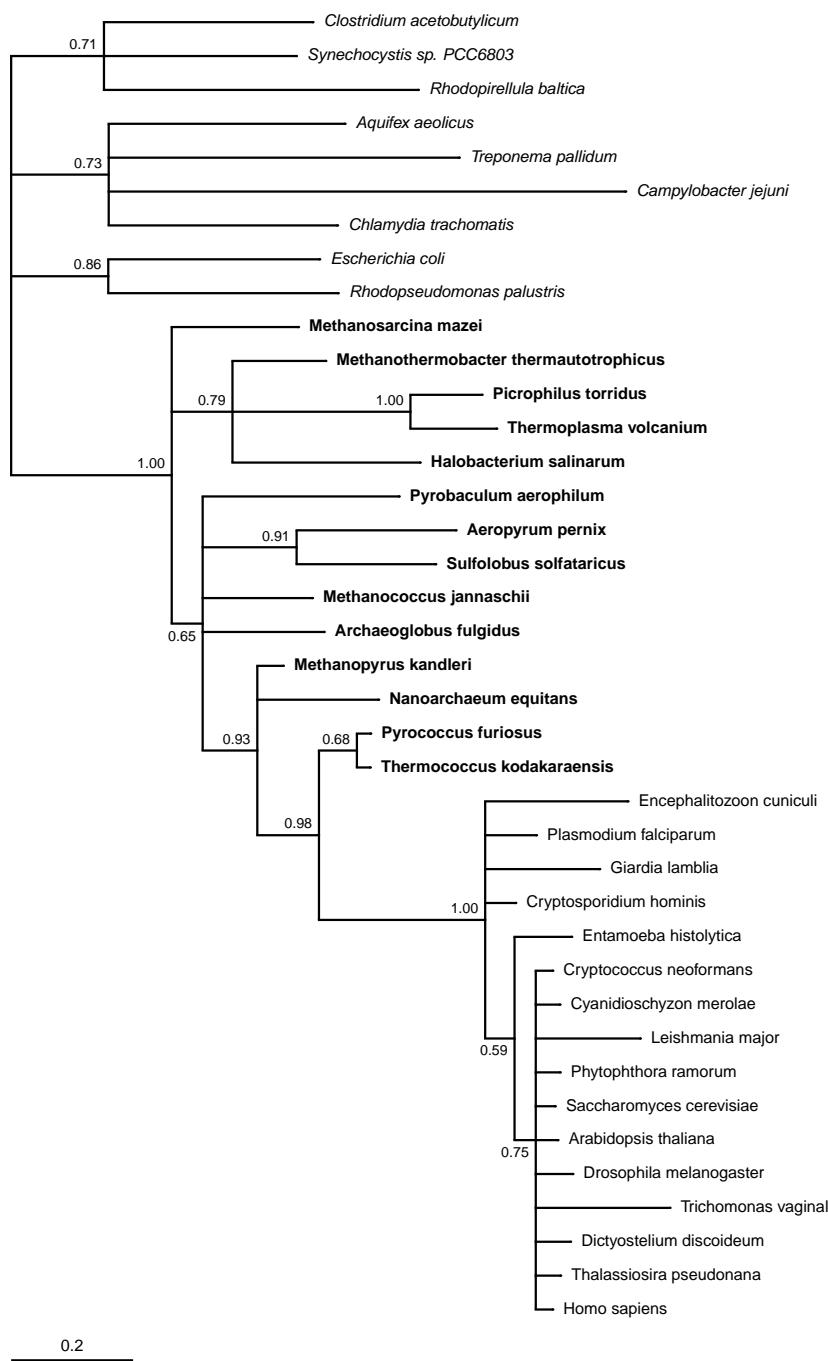
**Fig. S41:** Methionyl aminopeptidase – nTax = 39, nChar = 111 Substitution model: WAG+I+Γ+4CV Composition homogeneity test P value = 0.0618



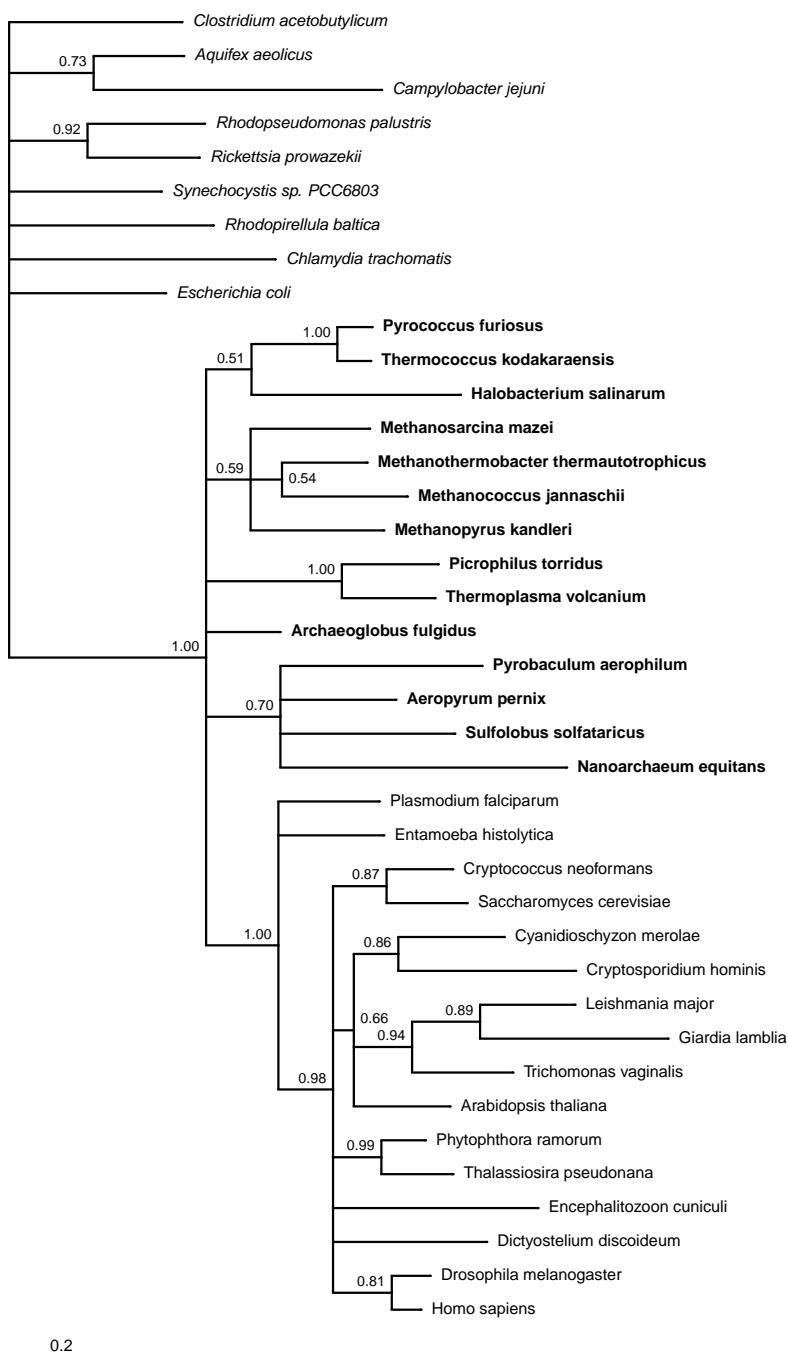
**Fig. S42:** Protein transport protein Sec61 $\alpha$  – nTax = 40, nChar = 102 Substitution model: WAG+Γ+1CV  
Composition homogeneity test P value = 0.3359



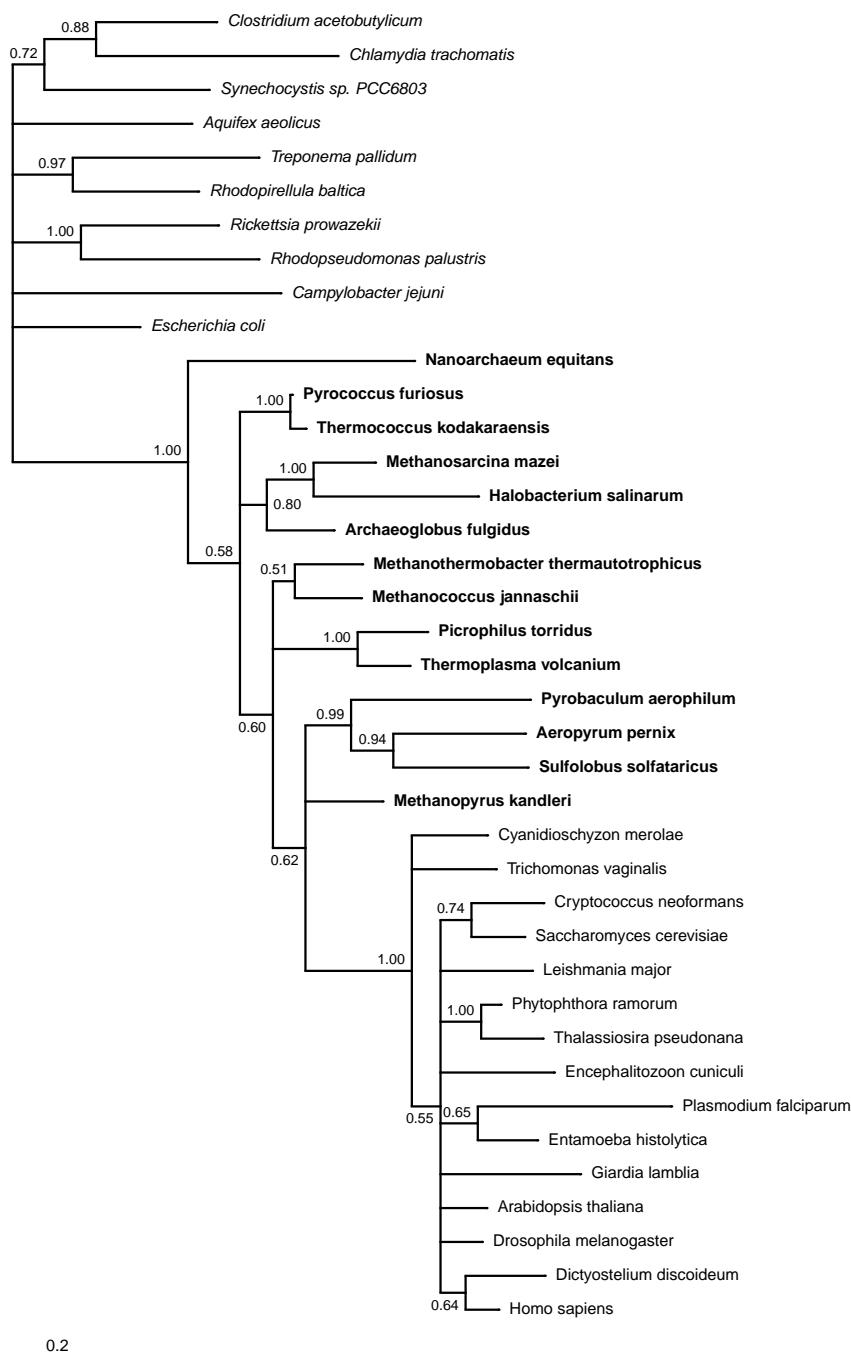
**Fig. S43:** Transitional endoplasmic reticulum ATPase – nTax = 40, nChar = 213 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.0861



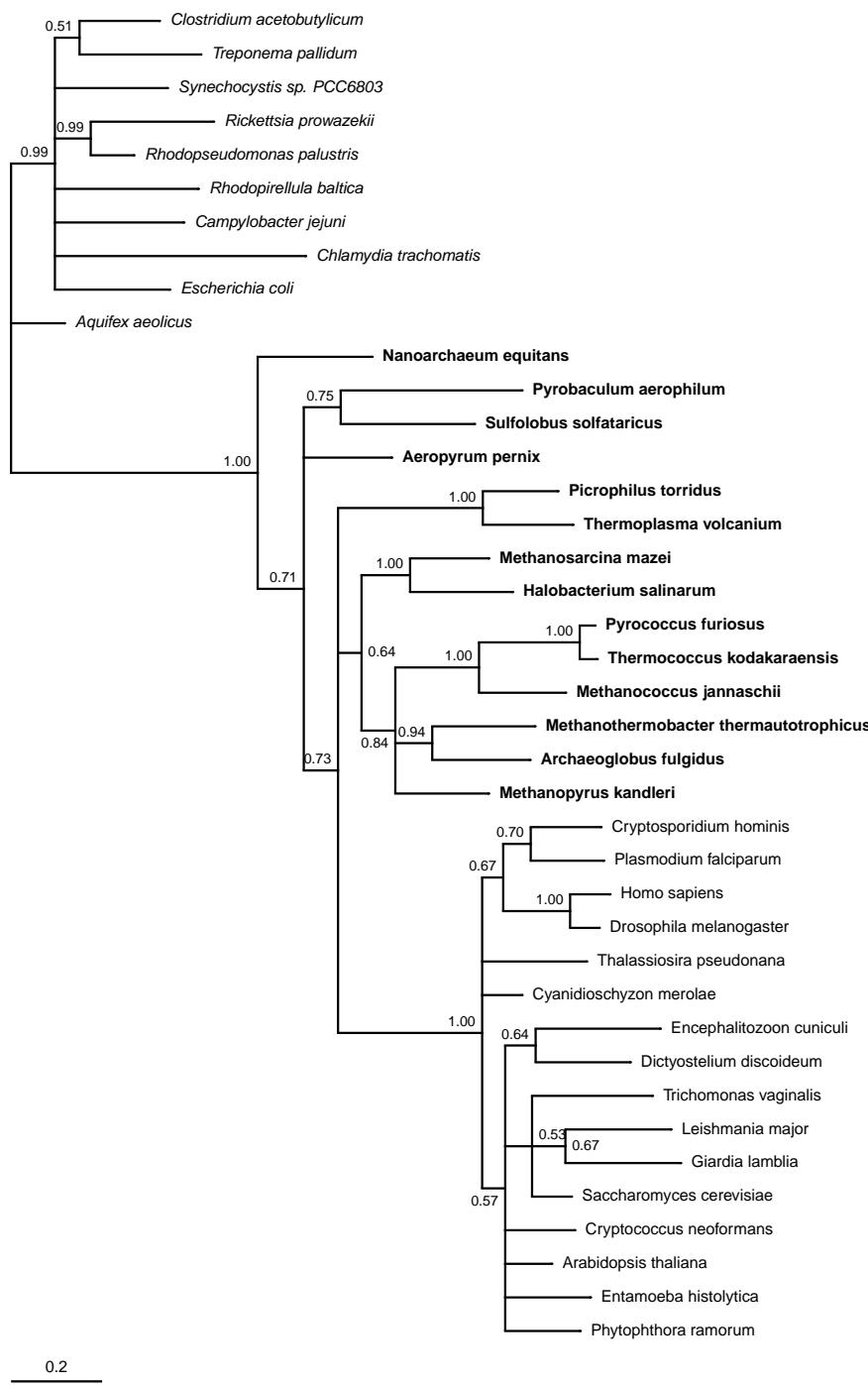
**Fig. S44:** Pseudouridine syntase component dyskerin – nTax = 39, nChar = 111 Substitution model: WAG+I+Γ+2CV Composition homogeneity test P value = 0.1130



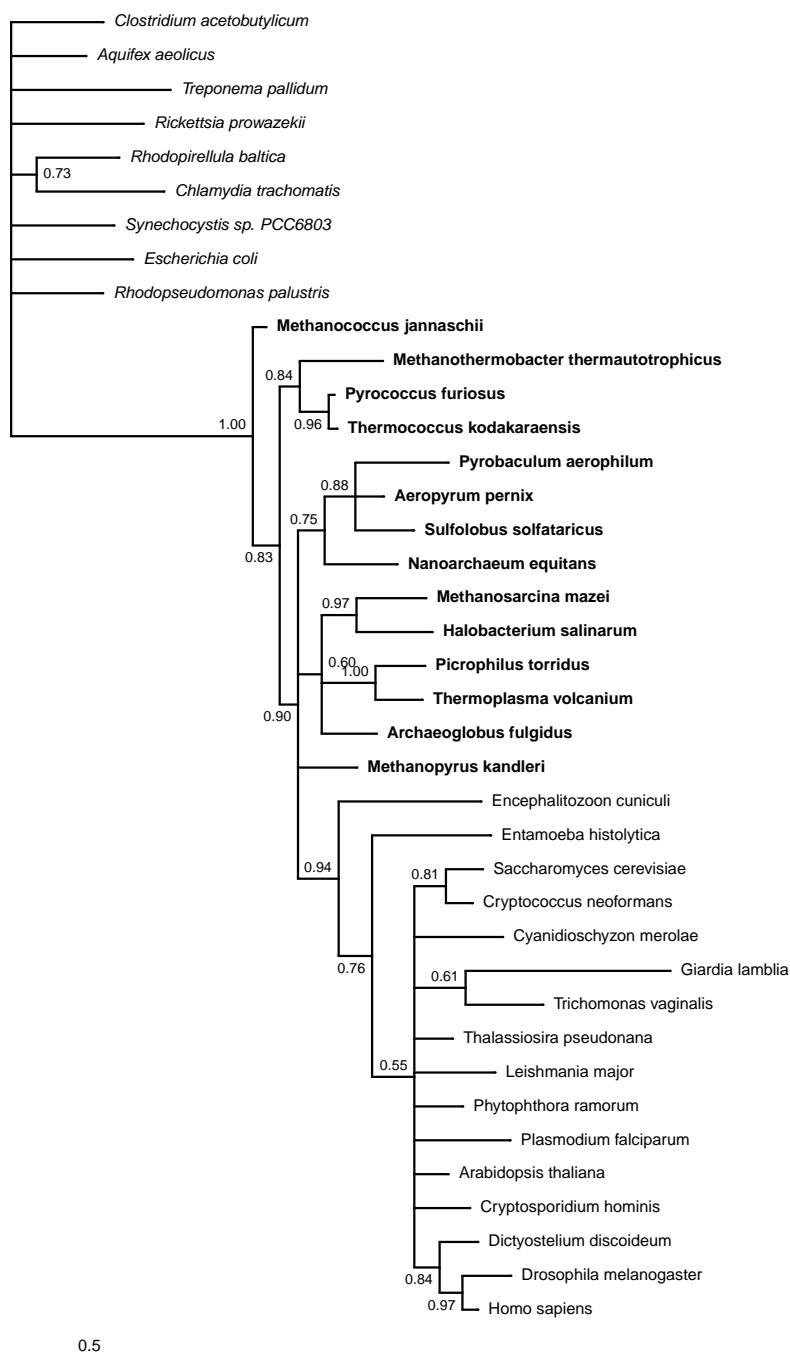
**Fig. S45:** Phenylalanine-tRNA ligase ( $\beta$ ) – nTax = 39, nChar = 97 Substitution model: WAG+I+ $\Gamma$ +9CV  
Composition homogeneity test P value = 0.0550



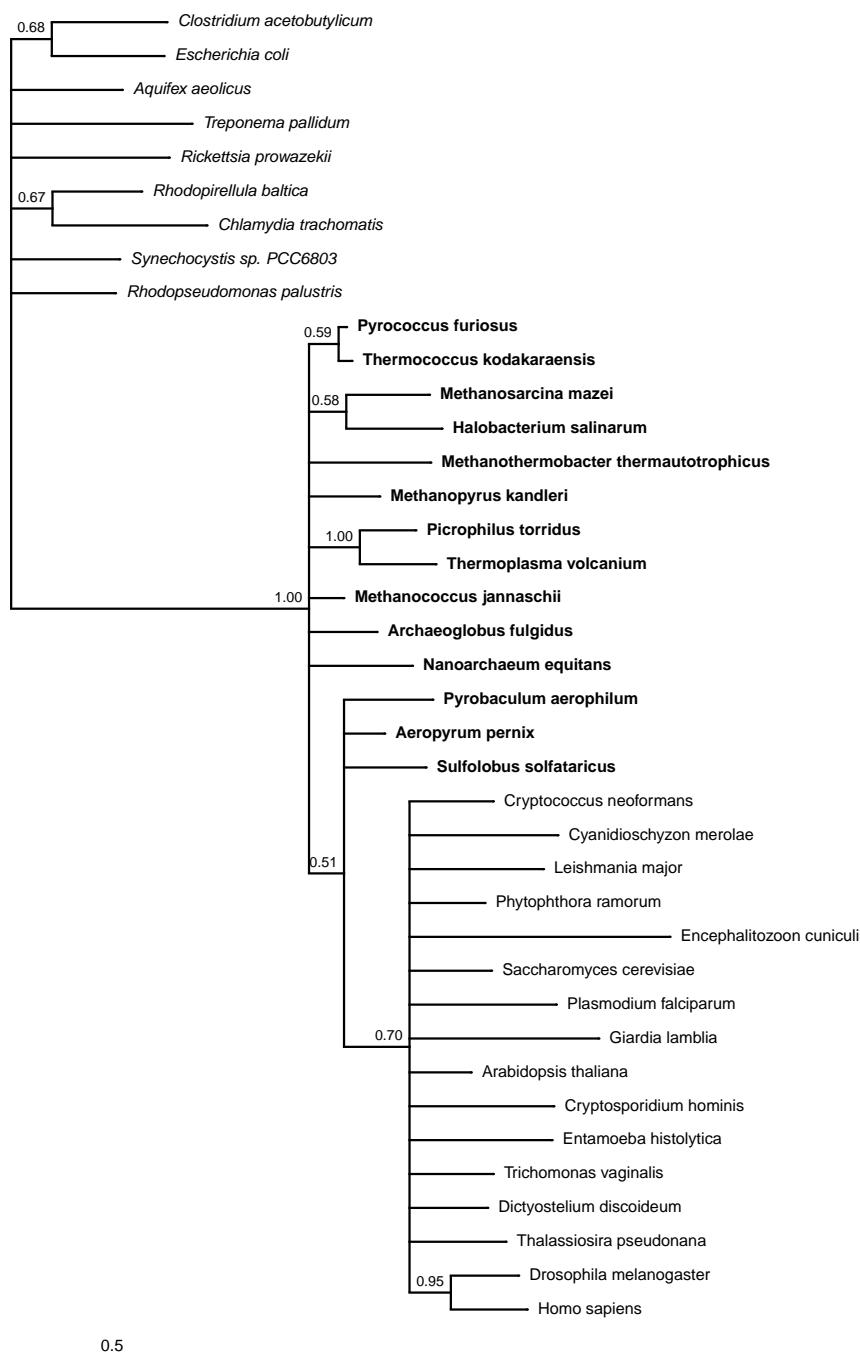
**Fig. S46:** O-sialoglycoprotein endopeptidase – nTax = 39, nChar = 176 Substitution model: WAG+I+Γ+6CV  
Composition homogeneity test P value = 0.0772



**Fig. S47:** Translation initiation factor IF-2 – nTax = 40, nChar = 187 Substitution model: WAG+I+G+4CV  
Composition homogeneity test P value = 0.0512



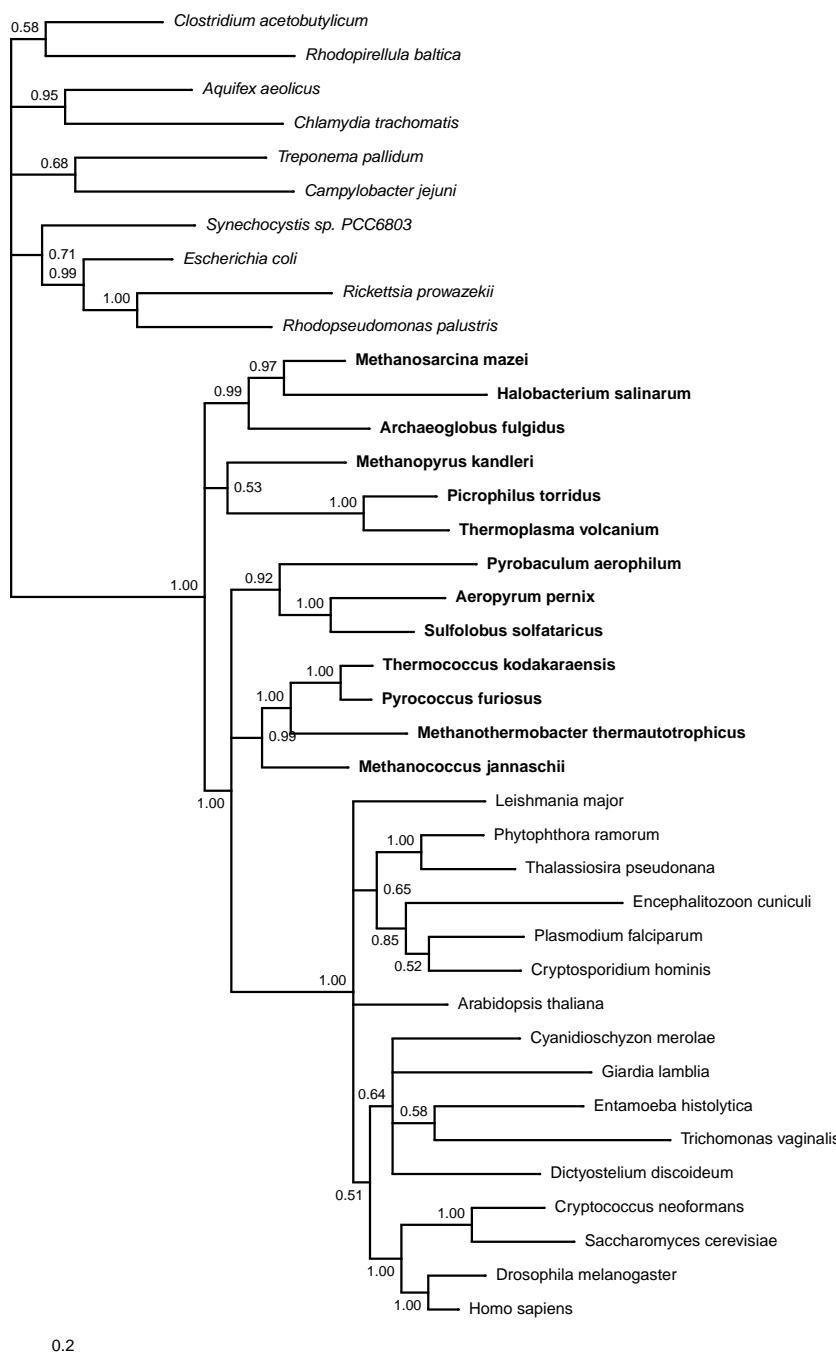
**Fig. S48:** Replication factor C subunit 2 – nTax = 39, nChar = 141 Substitution model: WAG+I+Γ+2CV  
Composition homogeneity test P value = 0.0766



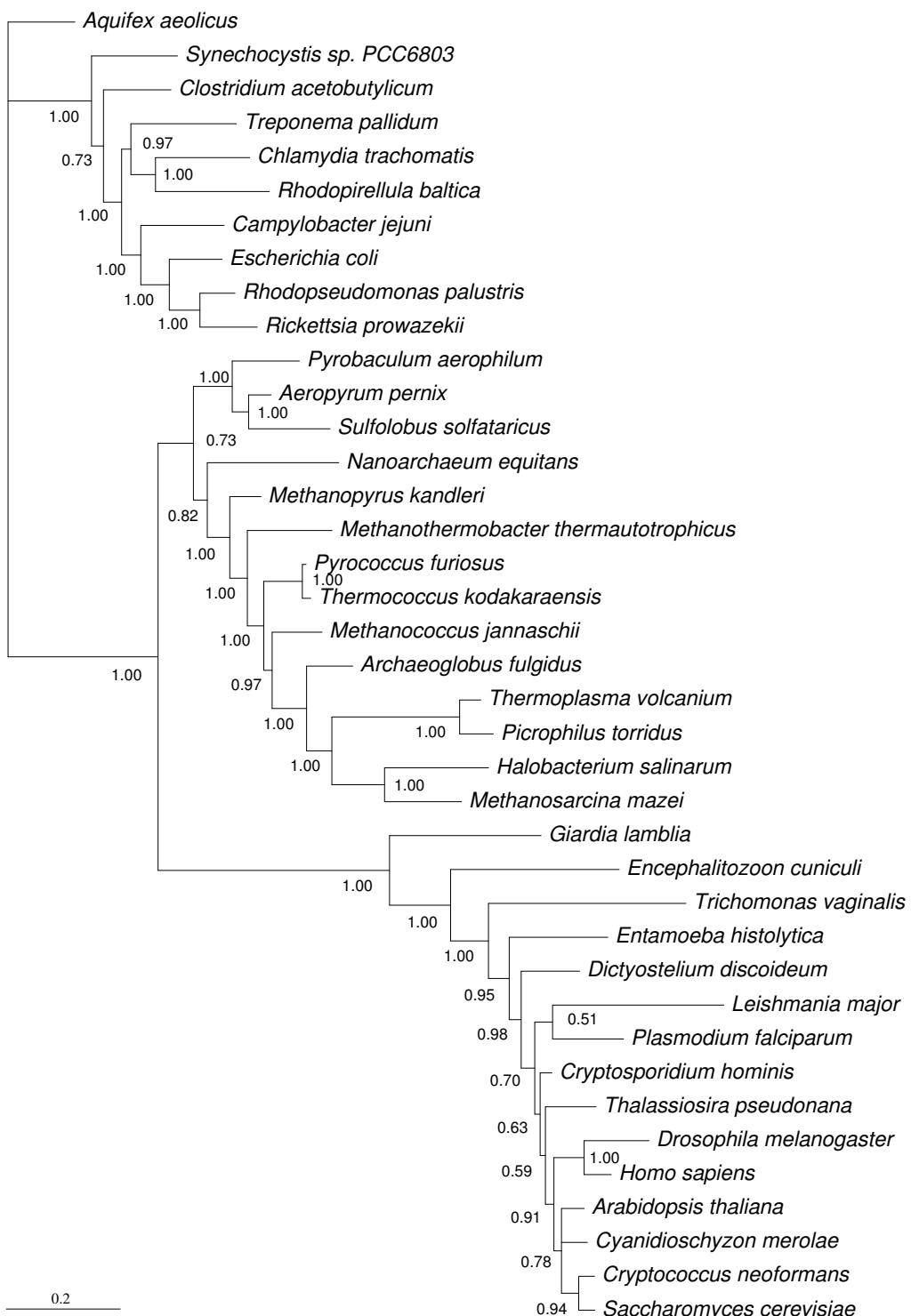
**Fig. S49:** Replication factor C subunit 4 – nTax = 39, nChar = 119 Substitution model: WAG+I+Γ+2CV  
Composition homogeneity test P value = 0.2844



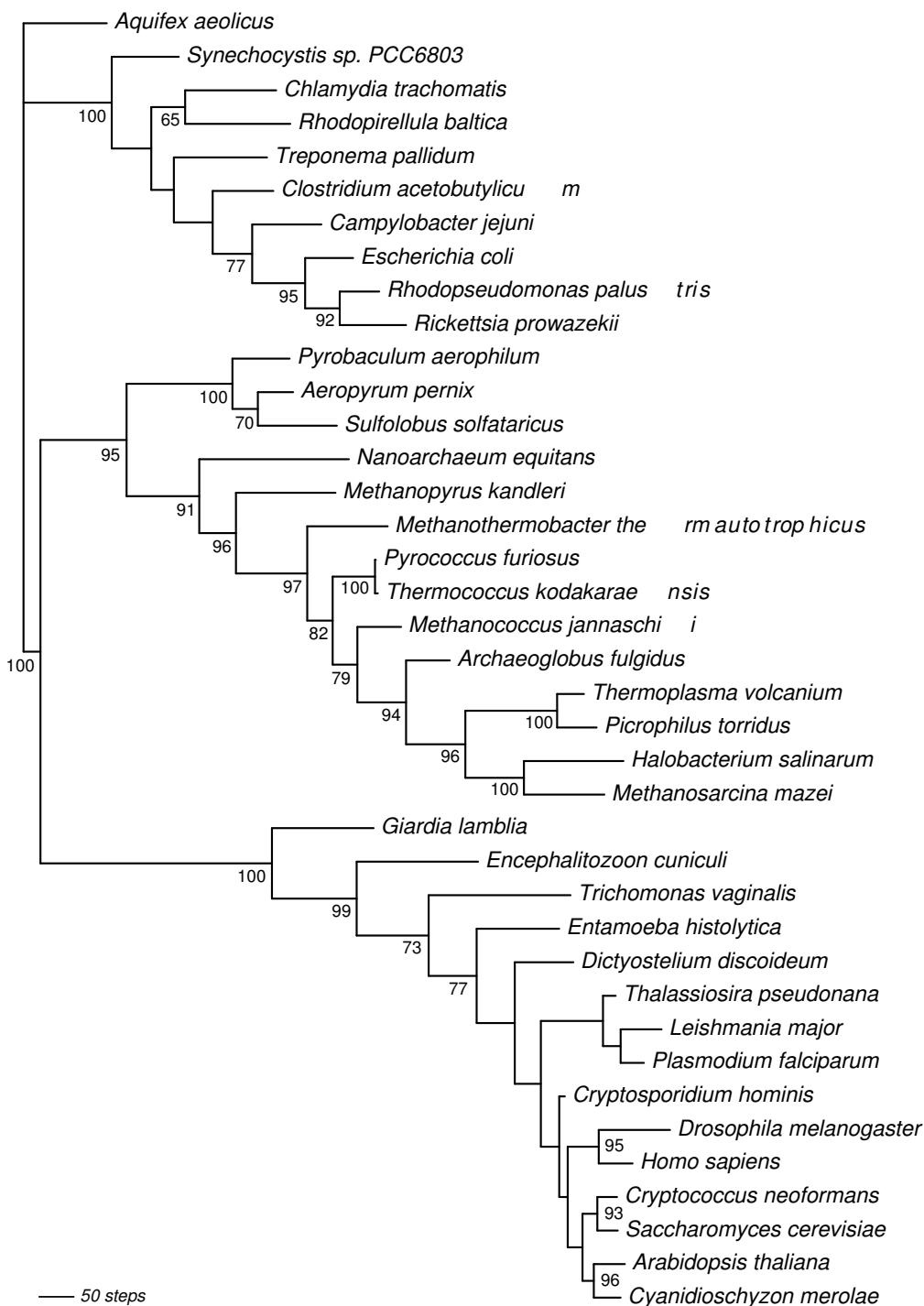
**Fig. S50:** Signal recognition particle receptor alpha subunit (SR- $\alpha$ ) – nTax = 39, nChar = 159 Substitution model: WAG+I+Γ+4CV Composition homogeneity test P value = 0.0659



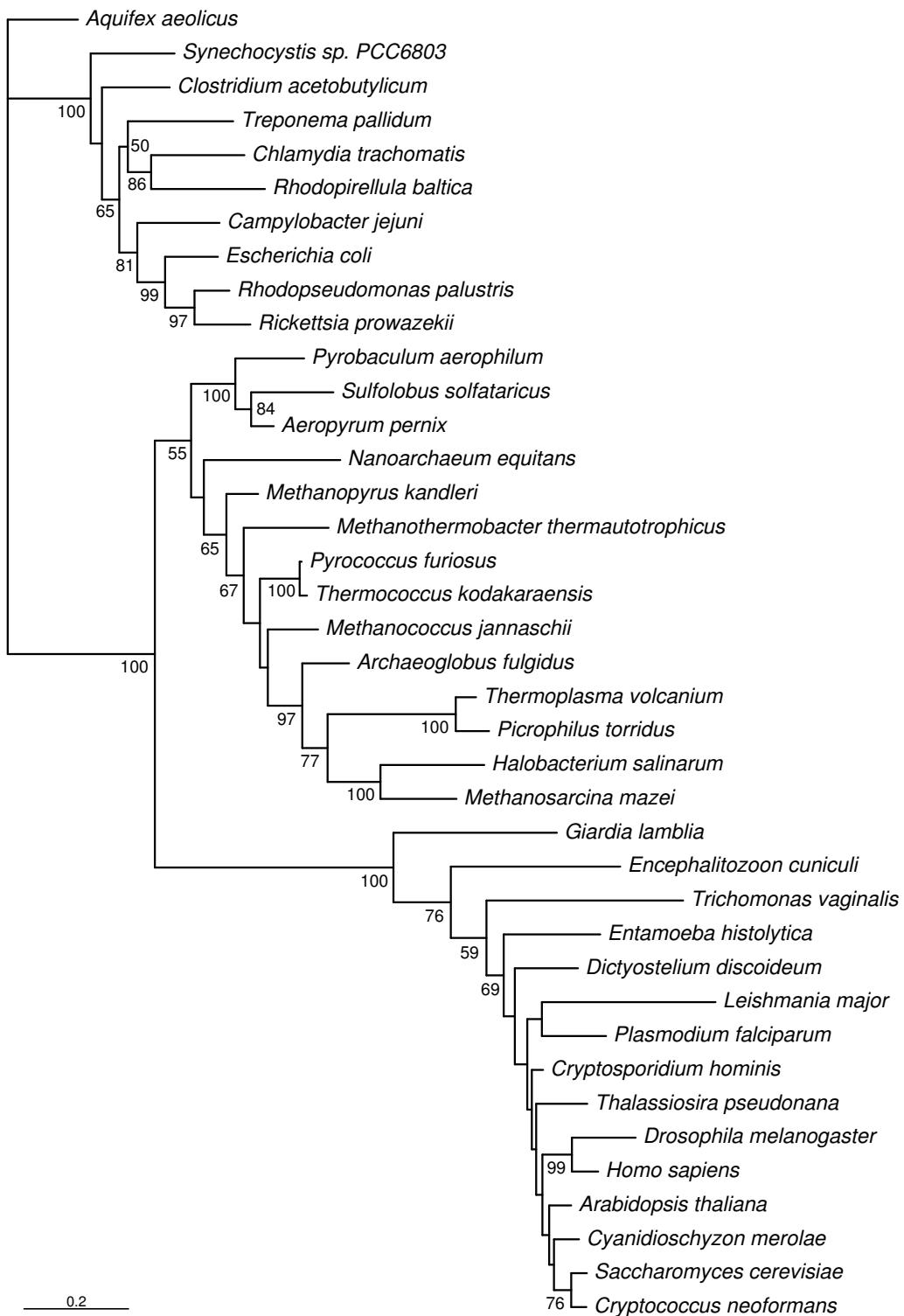
**Fig. S51:** Signal recognition particle recognition component (SRP54) – nTax = 39, nChar = 243 Substitution model: WAG+I+Γ+9CV Composition homogeneity test P value = 0.0342



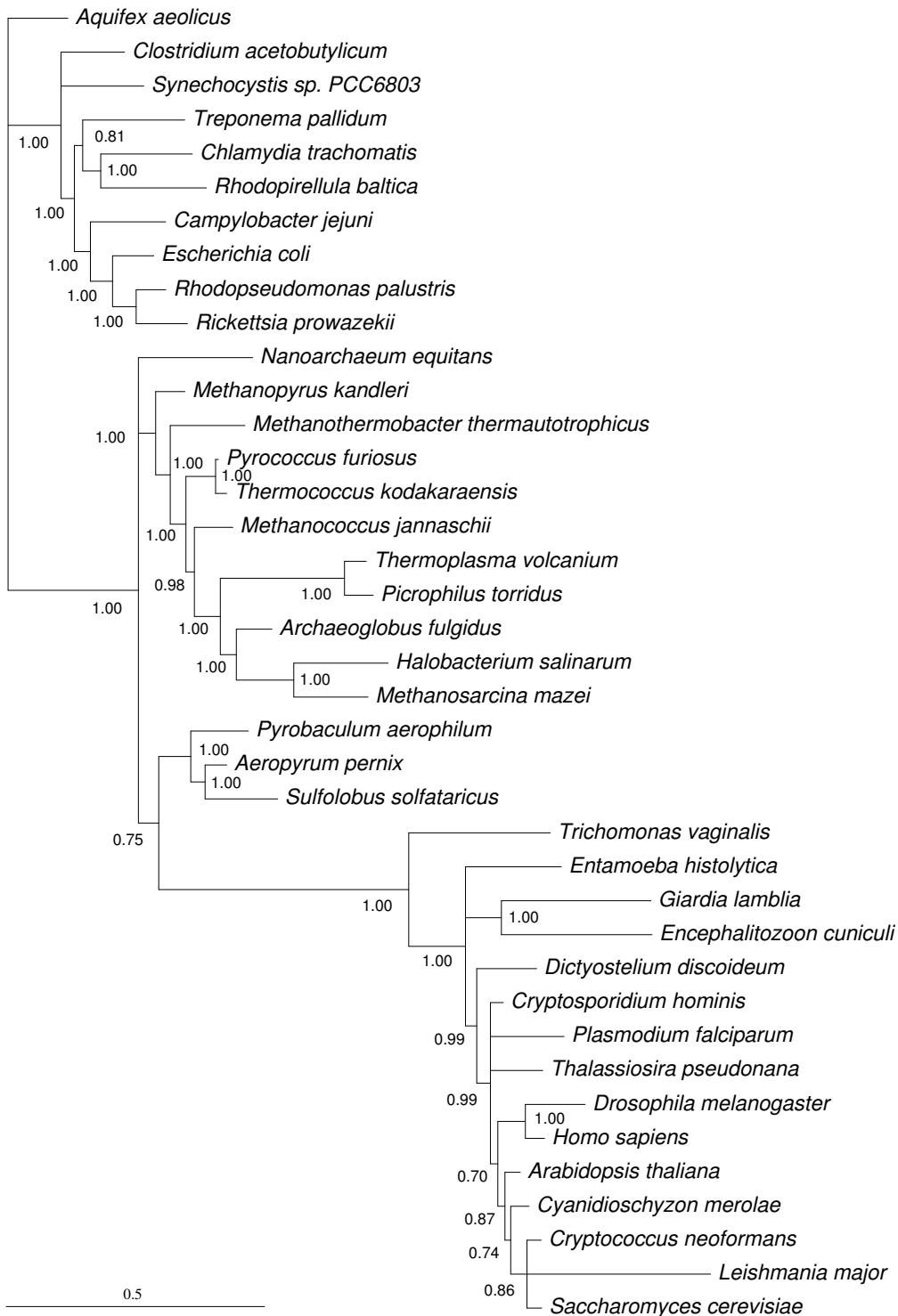
**Fig. S52:** Combined SSU & LSU rRNA: consensus tree of 16,000 trees obtained from the posterior distribution of two MCMC analyses (MrBayes) with homogeneous composition across the tree [(GTR+ $\Gamma$ )x2].  $\log_e(L_m) = -23960.00$



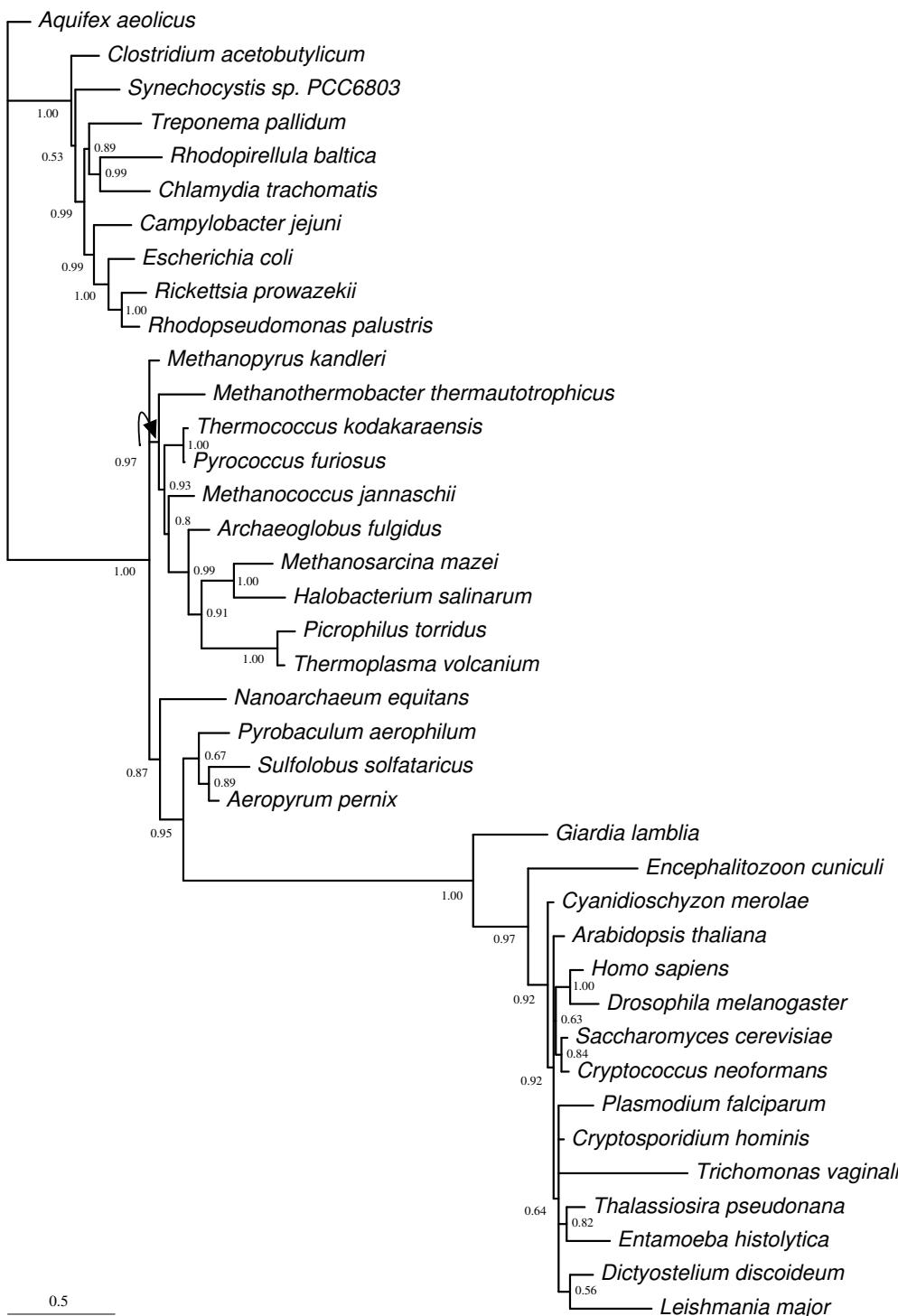
**Fig. S53:** Combined SSU & LSU rRNA: one of two equally-weighted maximum parsimony trees (length: 5498 steps) with maximum parsimony bootstrap (300 replicates) proportions ( $\geq 50\%$ ) indicated at nodes. Branch lengths were calculated using ACCTRAN optimisation. The second MP tree differs from that shown by placing *Clostridium acetobutylicum* sister to a clade including all eubacteria except *Aquiflex aeolicus* and *Synechocystis* sp.



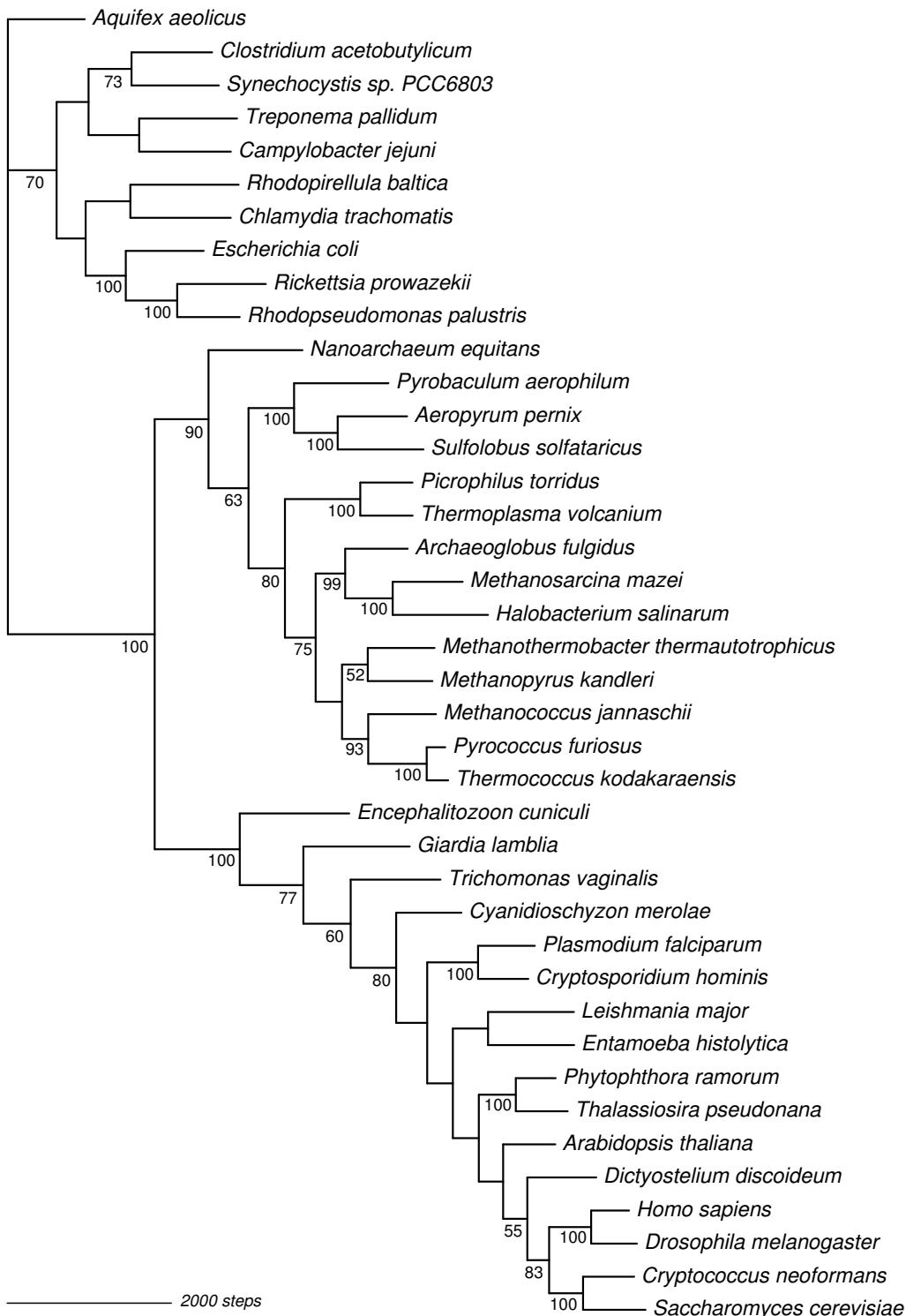
**Fig. S54:** Combined SSU & LSU rRNA: optimal maximum likelihood tree with bootstrap proportions ( $\geq 50\%$ ) indicated at nodes (GTR+ $\Gamma$ , 100 replicates).  $L_n = -23906.11$



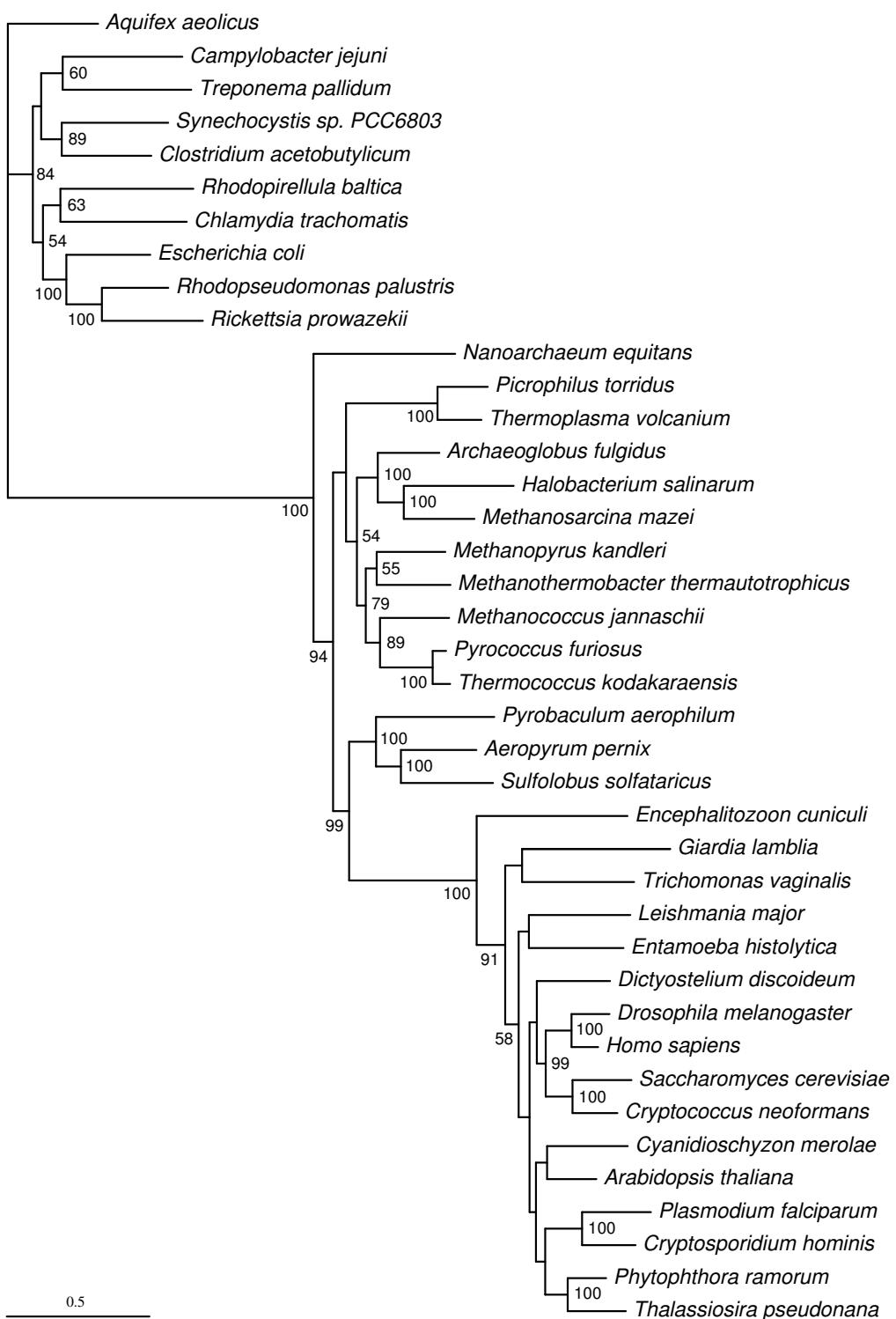
**Fig. S55:** Combined SSU & LSU rRNA: consensus tree of 10,000 trees obtained from the posterior distribution of an MCMC analyses with heterogeneous composition across the tree [(GTR+Γ+2CV)×2].  $\log_e(L_m)$  = -23507.36 Posterior predictive simulations of  $X^2$  - SSU: original stat = 468.06, P = 0.3810 (range = 186.57-1014.93, mean = 449.61), LSU: original stat = 759.69, P = 0.7515 (range = 475.46-1589.75, mean = 845.98). By contrast, posterior predictive simulations of  $X^2$  for the homogeneous model; SSU: range = 29.55-210.79, mean = 70.49, and LSU: range = 27.03-151.45, mean = 63.13



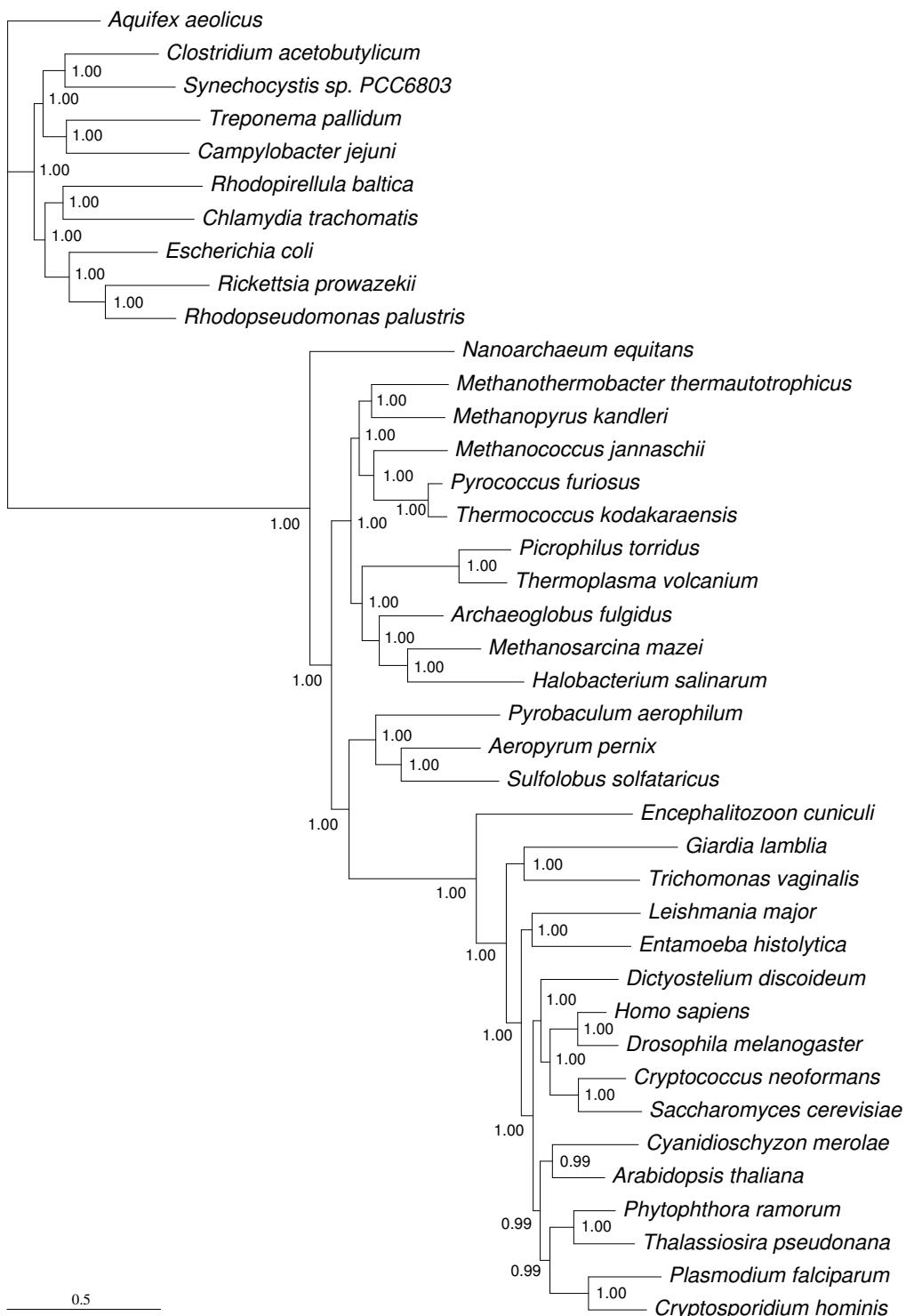
**Fig. S56:** Combined SSU & LSU rRNA: consensus tree of 21,755 trees obtained from the combined posterior distributions of four MCMC analyses with heterogeneous composition across the data using the CAT model (+Γ).  $\log_e(L_m) = -21916.59$



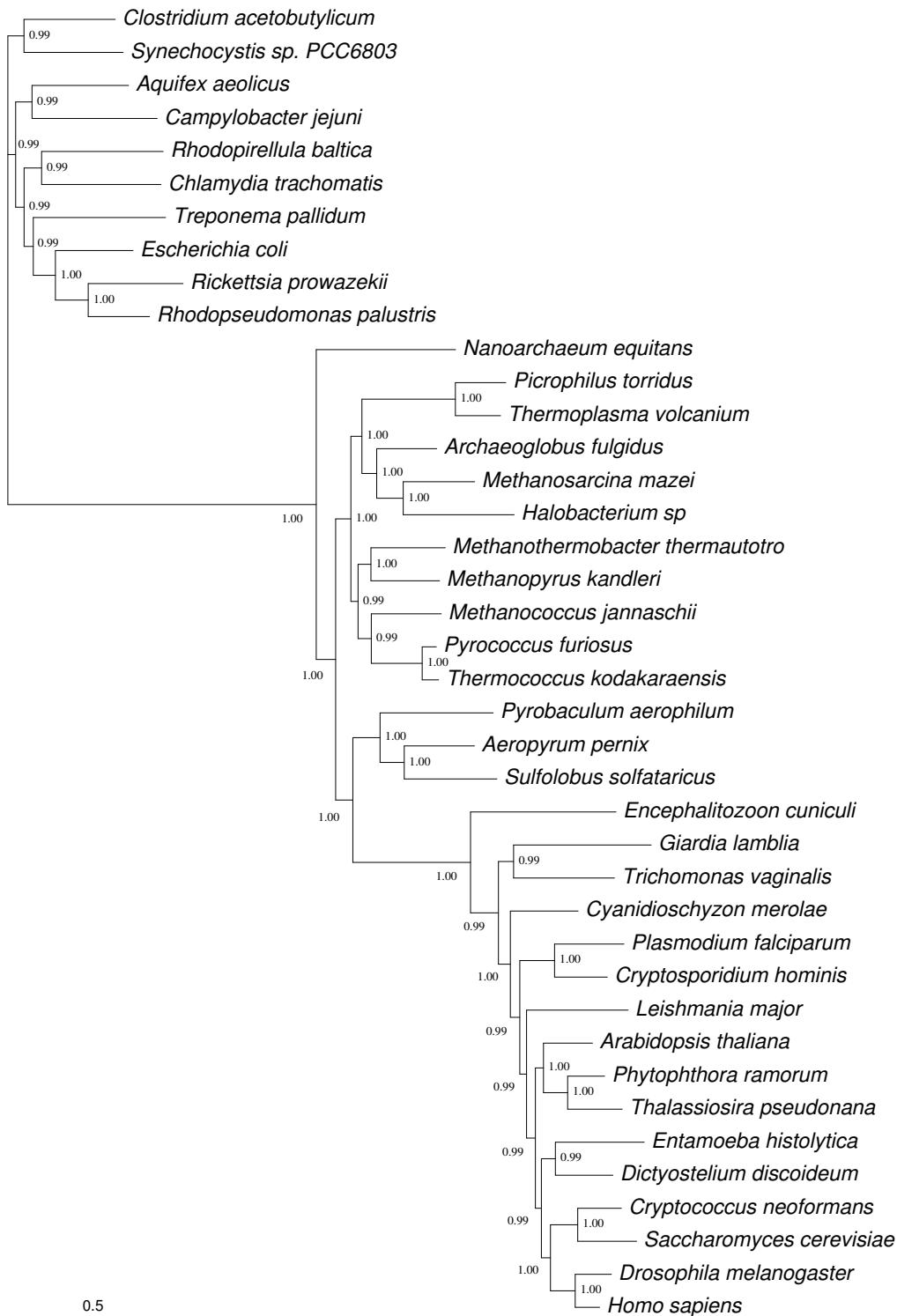
**Fig. S57:** Combined 45 proteins (standard amino-acid coding): one of two equally-weighted maximum parsimony trees (length: 56751 steps) with bootstrap (300 replicates) proportions ( $\geq 50\%$ ) indicated at nodes. Branch lengths were calculated using ACCTRAN optimisation. The second MP tree differs from that shown by placing *Methanopyrus kandleri* as the sister-group to a clade consisting of *Methanothermobacter thermautotrophicus*, *Archaeoglobus fulgidus*, *Methanosaeca mazaei*, and *Halobacterium salinarum*.



**Fig. S58:** Combined 45 proteins (standard amino-acid coding): the optimal maximum likelihood tree (WAG+Γ) with bootstrap (100 replicates) proportions ( $\geq 50\%$ ) indicated at nodes ( $L_n: -279967.06$ ).



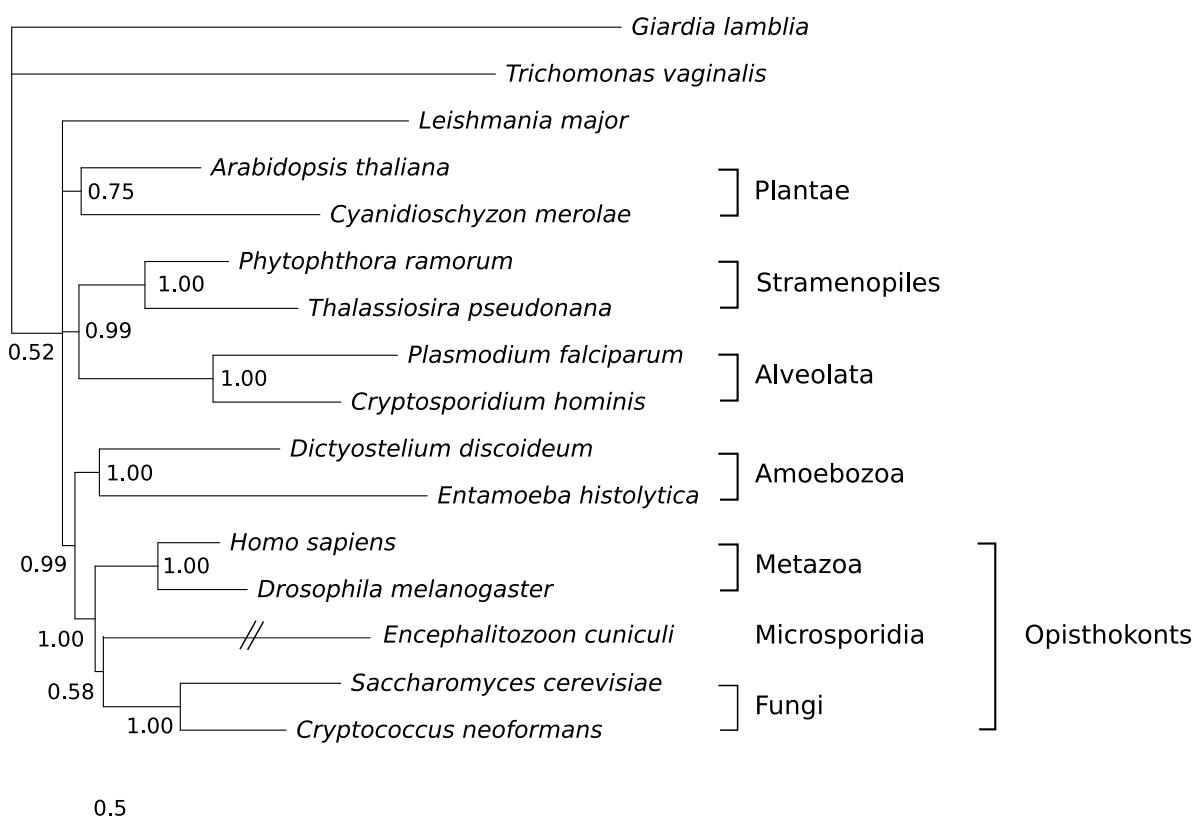
**Fig. S59:** Combined 45 proteins (standard amino-acid coding): consensus tree of 2,000 trees obtained from the posterior distribution of an MCMC analysis (MrBayes) with homogeneous composition across the data and tree (WAG+Γ).  $\log_e(L_m) = -280750.66$



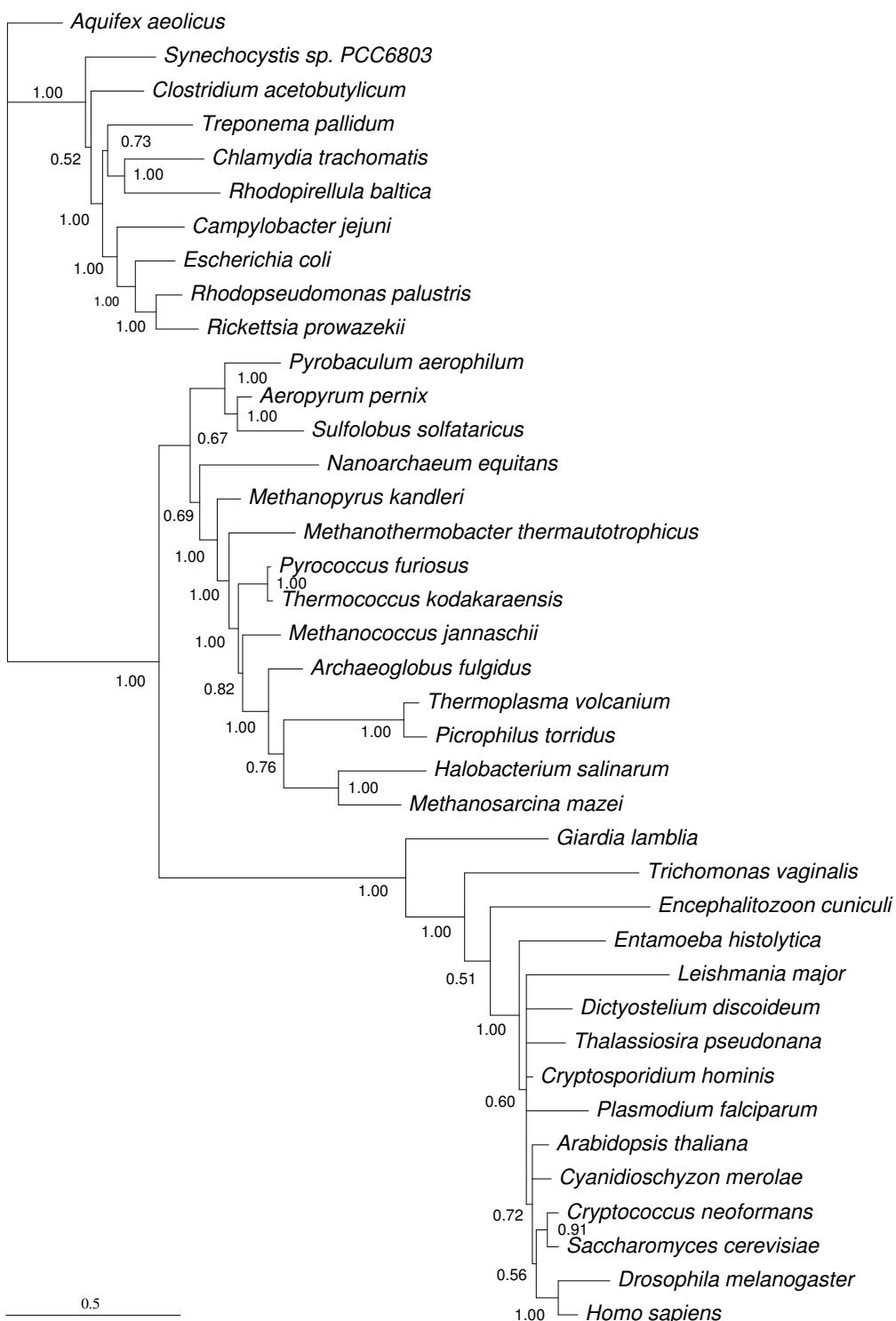
**Fig. S6o:** Combined 45 proteins (standard amino-acid coding): consensus tree of 2,162 trees obtained from the posterior distribution of an MCMC analyses with heterogeneous composition (WAG+Γ+26CV).  $\log_e(L_m) = -277039.32$  Posterior predictive simulations of  $X^2$  original stat 4266.98, range = 2273.71-3798.63, mean = 2925, P = 0.0000. By contrast, posterior predictive simulations of  $X^2$  for the homogeneous model: range = 331.69-551.24, mean = 422.57



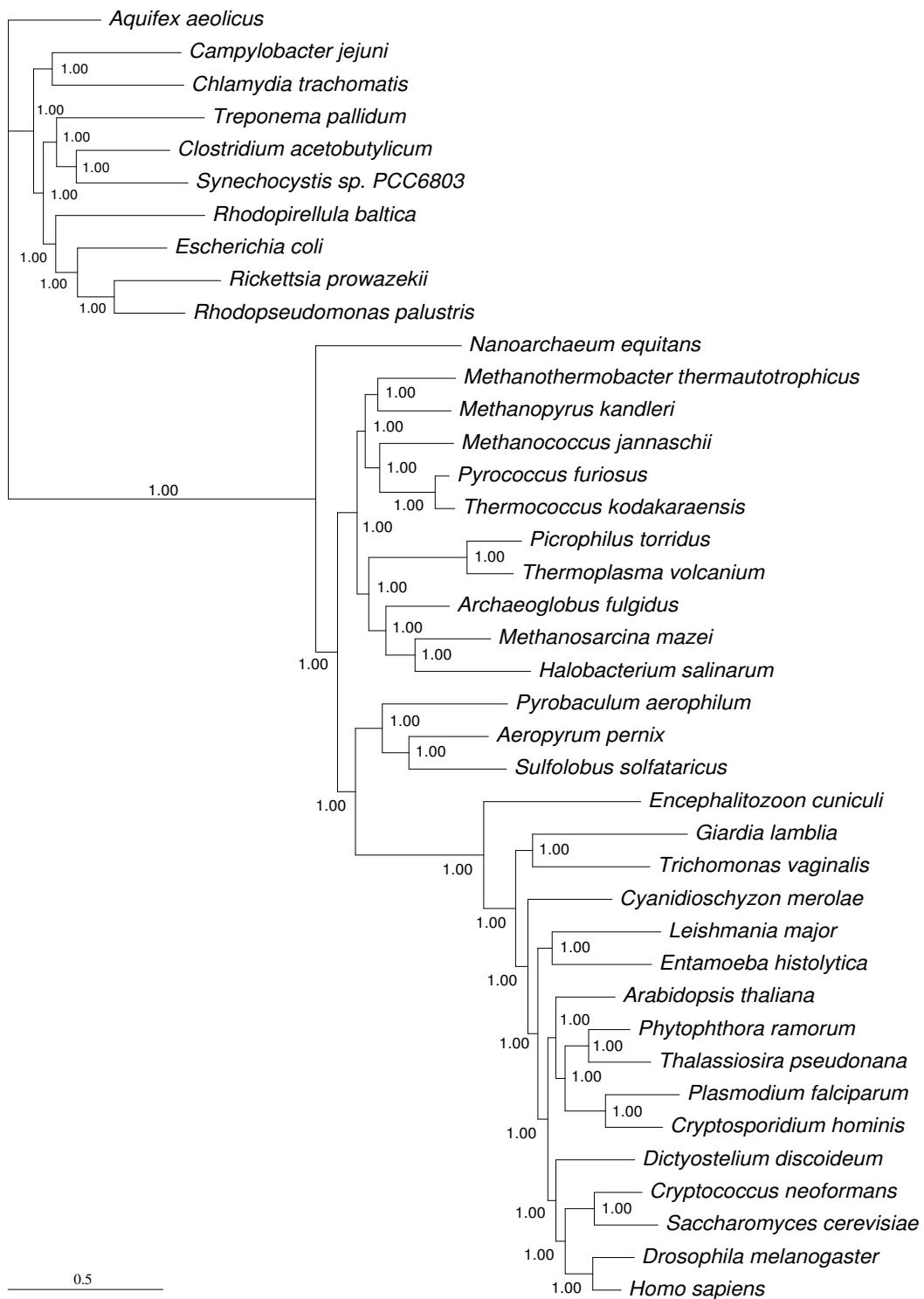
**Fig. S61:** Combined 45 proteins (Dayhoff recoded data): consensus tree of 8,544 trees obtained from the posterior distribution of an MCMC with the CAT model using Dayhoff recoded data.  $\log_e(L_m) = -112345.49$ , mode number of categories ( $k$ ) = 93.59



**Fig. S62:** MCMC phylogenetic analysis of the eukaryote taxa alone with the CAT (+Γ) model. Scale bar represents 0.5 substitutions per site. Taxonomy follows Embley and Martin (7). The branch leading to *Encephalitozoon cuniculi* has been arbitrarily reduced in length; its true length is equivalent to 2.0 substitutions per site.



**Fig. S63:** Combined SSU & LSU rRNA: consensus tree of 16,000 trees obtained from the posterior distribution of an MCMC analyses with the covarion parameter [(GTR+ $\Gamma$ )x2+COV].  $\log_e(L_m) = -23637.60$



**Fig. S64:** Combined 45 proteins (standard amino-acid coding): consensus tree of 9,000 trees obtained from the posterior distribution of an MCMC analyses with the covarion parameter [(WAG+ $\Gamma$ )x2+COV].  $\log_e(L_m) = -280850.21$